

# EXPLANATION IN CONTEXTS OF CAUSAL COMPLEXITY

by

**Lauren N. Ross**

B.S., California Polytechnic State University, San Luis Obispo, 2007

M.D., University of California, Irvine, 2011

Submitted to the Graduate Faculty of  
the Kenneth P. Dietrich School of Arts and Sciences in partial  
fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH  
DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Lauren N. Ross

It was defended on

April 14th 2016

and approved by

Robert Batterman, Philosophy Department

Mazviita Chirimuuta, HPS Department

William C. Wimsatt, Philosophy Department, University of Chicago

Dissertation Directors: James F. Woodward, HPS Department

Kenneth F. Schaffner, HPS Department

# **EXPLANATION IN CONTEXTS OF CAUSAL COMPLEXITY**

Lauren N. Ross, PhD

University of Pittsburgh, 2016

My dissertation examines common types of causal complexity in the biological sciences, the challenges they pose for explanation, and how scientists overcome these challenges. I provide a novel distinction between two types of causal complexity and I analyze explanatory patterns that arise in these contexts. My analysis reveals how explanation in the biological sciences is more diverse than mainstream accounts suggest, which view most or all explanations in this domain as mechanistic. I examine explanations that appeal to causal pathways, dynamical models, and monocausal factors and I show how these explanations are guided by considerations that have been overlooked in the extant literature. My project explores connections between these explanatory patterns and other topics of interest in philosophy and general philosophy of science, including: reduction, multiple realizability, causal selection, and the role of pragmatics in explanation.

## TABLE OF CONTENTS

<b>PREFACE</b> . . . . .	viii
<b>1.0 INTRODUCTION</b> . . . . .	1
<b>2.0 CAUSAL COMPLEXITY IN PSYCHIATRIC GENETICS</b> . . . . .	5
2.1 Introduction . . . . .	5
2.2 Background and motivation . . . . .	8
2.3 Common causal pathways to disease . . . . .	11
2.3.1 Parkinson’s disease . . . . .	11
2.4 Interventionist Interpretation of Pathway Explanation . . . . .	14
2.4.1 The pathway concept: two cases of causal complexity . . . . .	16
2.4.2 Multicausality . . . . .	17
2.4.3 Causal heterogeneity . . . . .	19
2.4.4 Causal complexity and challenges for explanation . . . . .	23
2.5 Pathways and Mechanisms . . . . .	24
2.6 Conclusion . . . . .	27
<b>3.0 DYNAMICAL MODELS AND EXPLANATION IN NEUROSCIENCE</b> . . . . .	28
3.1 Introduction . . . . .	28
3.1.1 Dynamical models in neuroscience . . . . .	31
3.1.2 Explanatory Dynamical Model: the Canonical Model . . . . .	35
3.1.2.1 Reducing models of neural excitability . . . . .	37
3.1.2.2 Ermentrout-Kopell Theorem . . . . .	38
3.2 Analysis of the Canonical Model Approach . . . . .	42
3.2.1 Kaplan and Craver’s Mechanist Account . . . . .	43

3.2.2 Batterman’s Minimal Model Explanations . . . . .	45
3.3 Conclusion . . . . .	49
<b>4.0 RECONSIDERING THE MULTIPLE REALIZABILITY ARGUMENT AGAINST EXPLANATORY REDUCTION . . . . .</b>	<b>51</b>
4.1 Introduction . . . . .	51
4.2 The multiple realizably argument against reduction: both sides of the debate	53
4.2.1 The multiple realizability thesis . . . . .	53
4.2.2 Opposition to the multiple realizability argument against reduction . .	56
4.3 Analyzing the smoking example: multiple realizability and causal heterogeneity	59
4.3.1 Initial sketch of the problem . . . . .	60
4.3.2 Important features of the smoking example . . . . .	62
4.3.3 Smoking example: problems for appealing to the lower level carcinogens	65
4.3.4 Objections . . . . .	70
4.4 Causal heterogeneity: problems for reductive explanation . . . . .	73
4.4.1 Causal heterogeneity and the final common pathway strategy . . . . .	75
4.5 Conclusion . . . . .	81
<b>5.0 CAUSAL CONTROL: A RATIONALE FOR CAUSAL SELECTION . . . . .</b>	<b>82</b>
5.1 Introduction . . . . .	82
5.2 Disease traits and interventionist causes . . . . .	85
5.3 Selecting among interventionist causes . . . . .	90
5.3.1 Specific causes and causal control of broad scope . . . . .	91
5.3.2 Probable causal control . . . . .	93
5.3.3 Stable causal control . . . . .	96
5.4 Conclusion . . . . .	98
<b>6.0 CONCLUSION . . . . .</b>	<b>100</b>
<b>7.0 REFERENCES . . . . .</b>	<b>105</b>

## LIST OF FIGURES

1	Causal pathway from genotype to phenotype (Snustad and Simmons, 2012) . . . . .	10
2	Causal pathways implicated in Parkinson’s Disease . . . . .	13
3	X as a direct cause of Y . . . . .	15
4	Three causal paths to Y . . . . .	16
5	Common pathway in a situation of multicausality . . . . .	20
6	Craver’s representation of levels of mechanisms and constitutive relations between levels (Craver 2007, p. 189) . . . . .	25
7	Phase plane with vector field (Izhikevich (2007), p. 113) . . . . .	33
8	Graph of the frequency-current (F-I) relationship of class I and class II neurons (Izhikevich (2007), p. 14) . . . . .	36
9	Modeling techniques in neuroscience (Izhikevich (2006), p. 279) . . . . .	39
10	Physiological state diagram of a Class I neural system (Hoppensteadt & Izhike- vich (1997), p. 228) . . . . .	40
11	The solution $x(t)$ of (3.3) is mapped to the solution $\theta(t)$ of (3.4), the canonical model (Hoppensteadt & Izhikevich (1997), p. 119) . . . . .	41
12	Multiple realizability (Fodor, 1975) . . . . .	54
13	Synchronic (d) and diachronic (e) relations between lower level and higher level phenomena (Sober (1999), p. 544) . . . . .	57
14	This figure represents the (a) multiple realization and (b) causal heterogeneity found in the smoking example. Relationships of realizations are represented with dashed lines and relationships of causation are represented with solid arrows	64

15	Convergence of gene variants on common processes: Cannon and Keller’s “Watershed model of the pathway between upstream genes and downstream phenotypes” (Cannon and Keller (2006), p. 273) . . . . .	78
----	--	----

## PREFACE

I would like to thank my committee for all of the support and helpful feedback they have provided me throughout my work on this dissertation. In particular, I would like to thank Bill Wimsatt and Mazviita Chirimuuta for their comments on earlier drafts of this work and Katie Tabb and Zina Ward for their encouragement and editorial help. Many of my philosophical projects have been motivated and facilitated by the work and progress that other philosophers have made. Ken Schaffner's examination of the final common pathway concept, Bob Batterman's account of minimal model explanation, and Jim Woodward's account of causation and causal explanation have all had serious influences on my work and have helped me to make my own contributions to the literature. I would especially like to thank Jim Woodward for the time and effort he has put into discussions of my work and various philosophical topics, and for his encouragement and support.

Finally, I would like to thank my family for the love and support that they have shown me throughout my graduate studies in philosophy and academic pursuits in general. My parents and my sister have supported me throughout my pre-medical, medical, and doctoral studies. I cannot imagine how I would have achieved the goals I have set my sights on without them.



## 1.0 INTRODUCTION

This dissertation concerns explanation and causation predominantly in the areas of biology, neuroscience, and medicine. According to mainstream philosophical views, all or most of the explanations in these scientific fields are mechanistic ([Kaplan and Craver, 2011](#); [Craver, 2009a](#); [Kaplan, 2011](#)). Many accounts of mechanistic explanation are motivated by the view that the way that we understand and explain biological phenomena often involves decomposing biological systems into their lower-level parts, the interactions between these parts, and how they are organized together ([Craver and Tabery, 2015](#)). Mechanistic philosophers have claimed that for a biological phenomenon of interest there are different lower-levels or layers of mechanisms for that phenomenon, where causal mechanical interactions are constrained to these levels and related to other levels in a constitutive or part-whole manner ([Craver, 2009a](#)). In addition to this, many mechanistic accounts of explanation maintain that the quality of an explanation increases as more mechanistic detail is provided ([Kaplan, 2011](#); [Kaplan and Craver, 2011](#)).

These dominant mechanistic views have encouraged a lively philosophical debate about scientific explanation. A significant amount of recent philosophical work examines whether scientific explanation *has* to be mechanistic and whether there are non-mechanistic or non-causal explanations in these scientific fields ([Woodward, 2013](#); [Dupré, 2013](#); [Chirimuuta, 2013](#); [Batterman and Rice, 2014](#)). My research contributes to this debate and it is motivated, in part, by a number of puzzles associated with these mainstream mechanistic accounts of explanation. A first puzzle is that scientists in these domains do not always view the provision of more mechanistic detail as increasing the quality of their explanations, although doing so may serve other purposes. A second puzzle, which is perhaps more problematic, is that scientists in these fields do not always cite mechanisms in the explanations that they provide.

They sometimes cite single causal factors, as in monocausal explanations of disease, they cite causal pathways, which are often explicitly distinguished from the notion of a mechanism, and they cite dynamical models, which are type of mathematical model that do not meet the requirements of mechanistic models. These three cases represent explanatory patterns that are importantly different from mechanistic explanation. One way in which these patterns differ from mechanistic explanation, is that they cite very few factors, while abstracting from large amounts of other seemingly relevant detail. In my dissertation, I examine these explanatory patterns in order to clarify their structure, the rationale behind their structure, and why they are found in some contexts, while not in others.

An important feature of my analysis involves the view scientists rely on these explanatory patterns to overcome challenges that causal complexity can pose for explanatory practice. I provide a novel distinction between two types of causal complexity. First, a biological phenotype can be causally complex in the sense that an instance of the phenotype is produced by many interacting causal factors. I refer to this type of causal complexity as multicausality. Explaining phenotypes that are multicausal can be challenging because this often requires identifying numerous causal factors, clarifying how they interact with each other, and distinguishing them from explanatorily irrelevant factors. Second, a biological phenotype can be causally complex in the sense that distinct instances of the same phenotype can be produced by different combinations of causal factors. This is a form of causal complexity that I call causal heterogeneity. Explaining phenotypes that are causally heterogeneous is complicated by the fact that there is no identifiable cause or causal process that can be cited to explain all instances of the phenotype. These situations lack identifiable causal factors that “make a difference” to all instances of the outcome of interest ([Woodward, 2003](#)).

How do scientists overcome these challenges in their explanations of biological phenomena? What guides their identification of relevant and irrelevant explanatory detail? I address these questions by examining various explanatory patterns that arise in contexts of causal complexity and by providing an analysis of the rationale behind their use. One important consequence of my analysis is that it reveals how explanation in this domain is more diverse than mainstream philosophical accounts suggest, which view most or all of these explanations as mechanistic. I show how explanations that appeal to monocausal factors, causal pathways,

and dynamical models are used to circumvent challenges imposed by causal complexity and how they are not accommodated by dominant mechanistic accounts of explanation.

In chapter 2 I examine an explanatory pattern in neuropsychiatric genetics where the notion of a causal pathway is used to explain causally complex disease phenotypes. Scientists claim that these “pathway diseases” are best understood as genetic and environmental perturbations of common causal pathways leading to disease. I clarify three main features of this pathway concept and show how it facilitates explanation of disease phenotypes that are characterized by multicausality and causally heterogeneity. This work highlights differences between the notion of a causal “pathway” and the common philosophical notion of a causal “mechanism.” These differences help explain why mainstream mechanistic accounts of explanation are unable to accommodate the structure and rationale behind these types of “pathway” explanations.

Chapter 3 explores a form of explanation according to which dynamical models are used to explain universal neural firing behaviors. In this context, the universal behavior is a shared firing pattern exhibited by neural systems with different microstructural detail. Explaining these behaviors involves complications associated with causal heterogeneity, as the behavior to be explained is exhibited by systems with different lower-level mechanisms. Dynamical systems neuroscientists explain these behaviors by using mathematical abstraction techniques to reduce models of molecularly diverse systems to a single “canonical model.” As these abstraction techniques reduce models while preserving their qualitative behavior, they provide a principled means of explaining why microstructurally distinct systems exhibit the same universal behavior. I argue that this explanatory pattern shares similarities to Batterman’s (2002) account of minimal model explanation and that it refutes Kaplan and Craver’s (2011) claims that all explanations in this domain are mechanistic.

In chapter 4 I examine a relationship between causal heterogeneity and multiple realizability to argue for the limited nature of explanatory reduction in biology. A key feature of this analysis is the view that the particular challenges that multiple realizability poses for explanatory reduction are helpfully characterized in the context of causal explanation. In the beginning of this chapter I identify a significant problem with Sober’s (1999) purported refutation of the multiple realizability argument against reduction. I then examine case

studies from Sober’s analysis and from the biological sciences to specify under-appreciated challenges for reductive explanation. These examples help clarify constraints and motivations that guide which “level” of detail scientists cite in their explanations of biological phenomena.

In chapter 5 I examine the rationale behind causal selection in the context of disease explanation. Causal selection refers to a distinction between background conditions and “the” cause of some outcome of interest. A longstanding consensus in philosophy views causal selection as lacking any objective rationale and as guided, instead, by considerations that are arbitrary, pragmatic, and non-scientific. I argue against this position in the context of disease explanation. Disease causes are selected on the basis of the type of causal control they exhibit over a disease phenotype of interest. My analysis clarifies the principled rationale that guides this selection, and how it involves pragmatic and objective considerations that have been overlooked in the extant literature.

My dissertation clarifies particular explanatory patterns and strategies that scientists use to explain biological phenomena in contexts of causal complexity. It provides a positive account of the rationale that guides scientists’ identification of relevant and irrelevant explanatory detail and it explains why they use particular explanatory patterns in some situations, while not in others. An important result of this work is that it reveals the diversity of explanation in biology, in contrast to the mainstream view that all or most explanations in this domain are mechanistic. Finally, this project explores connections between specific explanatory patterns and other topics of interest in philosophy and general philosophy of science, including: reduction, multiple realizability, causal selection, and the role of pragmatics in explanation.

## 2.0 CAUSAL COMPLEXITY IN PSYCHIATRIC GENETICS

### 2.1 INTRODUCTION

Attempts to clarify the etiology of many clinically accepted psychiatric disorders seem to continually meet with unexpected challenges. Perhaps the most recent example of such a challenge was the inability of genome-wide association studies (GWAS) to identify the specific genomic sequences that were thought to be associated with, and causally relevant to, particular psychiatric disorders.<sup>1</sup> These and further studies continue to suggest that an extreme degree of genetic complexity underlies many of these conditions in at least two distinct senses. Some psychiatric disorders appear to result from (1) the combined effects of multiple gene variants in individual patients and (2) different gene variants among distinct patients with the same disorder. I refer to these two types of causal complexity as (1) multicausality and (2) causal heterogeneity, respectively. Multicausality is a form of token-level causal complexity where multiple gene variants work together in aggregate to produce a token instance of some disease phenotype of interest. In the scientific literature these gene variants are often said to be variants of “small effect,” in contrast to cases where single gene variants of “large effect” are responsible for the phenotype. Causal heterogeneity is a form of type-level of causal complexity where causal gene variants differ across population-wide cases of the disorder. In these cases scientists refer to such genes as “rare variants,” in contrast to “common variants,” because any given variant is responsible for only a portion of the population level disease. These types of complexity are not mutually exclusive and both can be found in the same disease phenotypes, in fact, both are found in schizophrenia and autism spectrum disorder (ASD) ([Geschwind, 2011](#), 411-412). Genetic causes provide

---

<sup>1</sup>([Visscher, Brown, McCarthy, and Yang, 2012](#)).

only a partial characterization of the causal complexity of these disorders. Extending our attention to environmental and other non-genetic causes, further supports the view that many psychiatric disorders involve an extreme degree of causal complexity.

Some view these types of causal complexity as evidence against the possibility of explaining psychiatric disorders. Among the reasons for this claim are that, first, while psychiatric conditions could (in principle) have causal explanations, these are prevented by current psychiatric classification, which fails to group patients on the basis of shared causal etiology. This is a common position in the philosophy of psychiatry literature and it often involves the claim that these clinically defined disorders are characterized by a degree of causal heterogeneity or “process heterogeneity” that renders attempts to identify some shared causal etiology impossible or highly unlikely (Poland, 2014, 34). Such claims involve the common assumption that disease explanation requires appealing to some shared causal etiology that is characteristic of and causally relevant to the disorder in question. This type of shared causal etiology is referred to as a “causal signature” (Murphy, 2014, 105), “disorder-specific pathophysiology” (Caspi and Moffitt, 2006, 586) or “shared causal process” (Zachar, 2014, 87). Although this is a common position in the philosophy of psychiatry literature, what is meant or required by shared causal etiology is often unspecified. A second claim denies the possibility of explaining mental or psychiatric disease phenotypes on the basis that such traits are multiply-realized by distinct lower-level details (Putnam, 1975; Fodor, 1974). This second point has been used to support the mind-brain dualist position, which maintains that mental phenomena cannot be explained by appealing to the lower-level details of the brain.

In order to better understand these disorders, psychiatric geneticists have recently employed techniques from systems biology with apparent success. These techniques have led to a “novel” explanation where the notion of a causal pathway is used to explain these disorders and other complex phenotypes that involve significant causal complexity.<sup>2</sup> This approach views such complex phenotypes as “pathway diseases,” which are best understood as genetic and environmental perturbations of common causal pathways leading to the disease

---

<sup>2</sup>(Luo, Huang, Jia, Li, Su, Zhao, and Gan, 2014, 39). For an informative discussion of the final common pathway concept see (Schaffner, 1998, 241), (Schaffner, 2008, 76). Aside from this work, the concept of a final common pathway has received relatively little attention in the philosophical literature.

phenotype of interest.<sup>3</sup> Although this work is ongoing, mainstream researchers advocate the use of this perspective in understanding schizophrenia, autism spectrum disorder (ASD), and bipolar disorder, among other psychiatric and non-psychiatric conditions.<sup>4</sup> Psychiatric geneticists view this pathway-perspective as providing a “novel explanation” for psychiatric disease phenotypes and their genetic heterogeneity.<sup>5</sup>

Growing support for this pathway perspective and its ability to accommodate the causal complexity associated with these disorders warrants further understanding this approach. What does it mean for researchers to call these pathway diseases? What type of novel explanation do they provide and how do they accommodate causal complexity? In this chapter, I focus on answering these questions by examining the structure of this explanatory pattern in psychiatric genetics. I will not argue that this explanatory pattern works for all psychiatric disorders or that future evidence will support this “disease-pathway hypothesis.” (Sullivan, 2012) My focus will be on current work in psychiatric genetics and those disorders that researchers find amenable to this pathway perspective. Key features of my analysis include an examination of this explanatory approach in the case of Parkinson’s disease and reliance on the interventionist account of causation (Woodward, 2003). I argue that the types of causal complexity associated with psychiatric disorders do not prevent explanation of these disease phenotypes, on the grounds commonly argued. That a disease phenotype involves either multicausality or causal heterogeneity does not indicate a lack of shared causal etiology or the irrelevance of lower level biological or neurological details. While these types of causal complexity can pose problems for explanation, I indicate that the pathway approach is a strategy that scientists use to circumvent these problems.

This chapter is structured as follows. The next section contains further discussion of the pathway concept and its relation to current techniques in systems biology. In the third section I examine the case of Parkinson’s disease as a model for the “pathway disease” explanatory pattern found in psychiatric genetics. Parkinson’s disease involves the same types of causal complexity that are found in many psychiatric disorders and the same appeal to the pathway concept. In the forth section I briefly describe Woodward’s interventionist account

---

<sup>3</sup>(Sullivan, 2012; Sullivan, Daly, and O’Donovan, 2012).

<sup>4</sup>(Sullivan, 2012; Martins-de Souza, 2012; Geschwind, 2008; Luo et al., 2014).

<sup>5</sup>(Luo et al., 2014, 39).

of causation, and use it to clarify how the pathway concept is used to provide explanations in situations of multicausality and causal complexity. At the end of this section I argue against common claims that these types of causal complexity prevent causal explanation of psychiatric disorders.

## 2.2 BACKGROUND AND MOTIVATION

Increased awareness of the genetic complexity of psychiatric disorders has encouraged the use of systems biology approaches to place these disorders within a broader biological context. A common feature of these approaches has been studying gene variants in terms of their downstream influences and within the context of whole networks of causal interactions between genes, proteins, cells, organ systems, and clinical phenotypes.<sup>6</sup> These approaches have led to the “disease pathway hypothesis,” which claims that the causal etiologies of psychiatric disorders are best understood at the “pathway level” (Sullivan, 2012). McClellan and King discuss this pathway perspective in understanding disease. As they state:

“The ultimate goal of gene discovery in complex disease is to identify and characterize biological pathways and processes critical to the disorder. Key pathways may be disrupted via many different causes—genetic, epigenetic, and environmental. Even if the illness in every affected individual arises from a different specific cause, each will nonetheless share disruption of related key biological processes. Defining the ways in which biological networks for common disease are impacted by mutation will contribute substantially to the understanding of their pathology and provide important targets for intervention.” (McClellan and King, 2010a, 216)

In another paper on schizophrenia as a pathway disease, Sullivan states:

The polygenicity of psychiatric disorders poses intriguing difficulties: how can these many genes be coherently tied together? A parsimonious hypothesis is that the polygenic basis of a psychiatric disorder is manifested in the regulation or function of one or more known or novel pathways. At present, we can only be vague about the meaning of the term ‘pathway’; current knowledge is limited, and the fundamental pathways could be biochemical, regulatory, developmental or anatomical and might correspond to a known pathway or a hitherto cryptic process. The idea arising from the observation of high polygenicity in schizophrenia is that the functional unit conferring risk may not be any single node but

---

<sup>6</sup>(Barabási and Oltvai, 2004).



rather the pathway itself. Thus, schizophrenia may be a cardinal example of a pathway disease. (Sullivan, 2012, 210)

These claims are motivated by recent research in psychiatric genetics, where a variety of techniques have been used to study biological and environmental influences on psychiatric disorders. This diversity has led to different uses of the pathway concept and interesting scientific discussions regarding its meaning and proper representation format.<sup>7</sup> I focus on a subset of this work where the concept of a pathway refers to a causal sequence, or set of causal steps, that outline some route of interest. This notion of a pathway is common in biology and inherent to examples such as developmental or neurodevelopment pathways, metabolic pathways, cell-signalling pathways, and gene expression pathways. In the cases I examine, this pathway concept is mainly used to refer to sequences of causal relationships from genetic and/or environmental factors to some phenotype of interest. This notion of a pathway from genotype to phenotype is represented in Figure 1 (Snustad and Simmons, 2012). When these pathways are discussed in the context of normal traits they involve highly complex causal interactions that lead to the trait of interest. However, when these pathways are discussed in relation to disease phenotypes, they are often represented in much less detail and viewed as far less complex. One reason for this, is that these pathways emphasize the specific pathogenic causes and their influences, while all other normal (causal and non-causal) factors are part of an assumed, but unrepresented background. These disease pathways outline some causal route by which a genetic or environmental insult cascades through a system, ultimately leading to the disease phenotype.

Elucidating or “tracing” these pathways is viewed as revealing information about the causal etiology of the disease in question. This is indicated by the fact that researchers refer to these pathways as “explanatory pathways” or “etiological pathways,” as they specify sequences of causal factors and relationships that explain their downstream effects.<sup>8</sup> The notion of a pathway from genotype to phenotype is also central to the endophenotype concept in psychiatric genetics. Gottesman and Shields defined endophenotypes as “intermediate phenotypes that form the causal links between genes and overt expression of disorders”

---

<sup>7</sup>(Lu, Sboner, Huang, Lu, Gianoulis, Yip, Kim, Montelione, and Gerstein, 2007; Papin, Price, Wiback, Fell, and Palsson, 2003).

<sup>8</sup> (Kendler, 2005, 5) (Kendler, 2013, 1060).

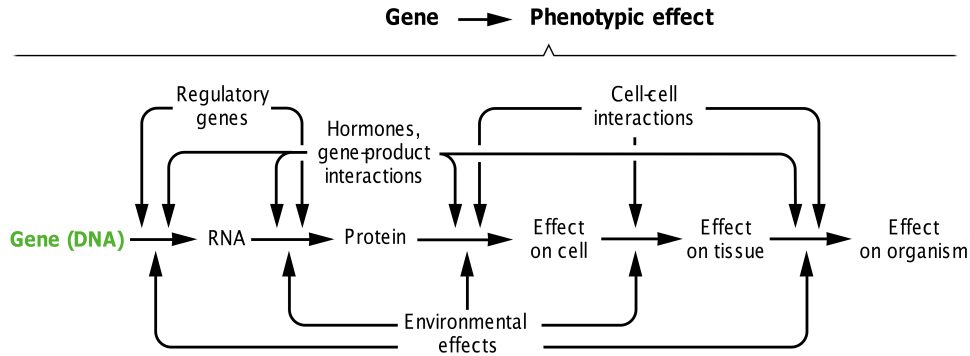


Figure 1: Causal pathway from genotype to phenotype (Snustad and Simmons, 2012)

and discussed them as a tool for elucidating the genetic causes of schizophrenia.<sup>9</sup> The identification of endophenotypes is said to be valuable, because their intermediate position along the causal pathway can be exploited to trace back (or upstream) to the genetic causes of the phenotype of interest. Tracing endophenotypes upstream of the disease phenotype allows for a way of “ ‘moving closer’ along the causal pathway to the ‘DNA level’ ” (Kendler and Neale, 2010, 789). Recent work in systems biology shares the goal of elucidating these causal pathways, but instead of working backward (upstream) from the phenotype, they focus on working forward (downstream) from gene variants. These studies start with genes variants that are associated with a given phenotype and study their direct effects to trace causal pathways from genes to phenotype. The influence of these gene variants is emphasized in this research because “gene expression is the first step on any pathway from genes to behavior” (Plomin, DeFries, Knopick, and Neiderhiser, 2012, 149). The influence of gene variants represents the first step along the pathways from genes to phenotype, while downstream portions of the pathway involve “higher levels,” such as cellular organization, neural synaptic connections, neural circuitry, and ultimately some higher level phenotype of interest.

<sup>9</sup> (Cannon and Keller, 2006, 268). Gottesman and Shields redescribed the endophenotype concept and discussed its importance in psychiatry, while John and Lewis first discussed the term in the context of insect research in 1966 (John and Lewis, 1966; Gottesman and Shields, 1972).

The use of this perspective in recent psychiatric genetics research had led to an interesting finding: although these disorders are associated with “seemingly disparate” gene variants, in some cases these factors all influence or converge on a shared causal pathway leading to disease (Yang, 2009, 421). In the context of discussing mental disorders, this is suggested by Kiesler who states that “A final common pathway denotes that heterogeneous causal factors that may operate in a particular disorder all converge on a shared link (physiological and/or psychological) that determines the core symptoms of the disorder” (Kiesler, 1999, 180). Researchers distinguish between common and distinct causal pathways, and view common causal pathways as providing an important type of common causal etiology. This distinction is discussed by Kendler, who states, “At one extreme, there may be dozens of biologically distinct pathways to illness with little or no sharing between them. At the other extreme—etiologic homogeneity—just one pathway to illness awaits discovery” (Kendler, 2013, 1060).

These claims motivate a number of questions. What does it mean for researchers to call these pathway diseases? Can appealing to common pathways explain causally complex diseases or is such talk merely figurative? Growing support for this perspective and its ability to accommodate causal complexity, motivates the need to gain a deeper understanding of this view.

## 2.3 COMMON CAUSAL PATHWAYS TO DISEASE

This section includes a short discussion of Parkinson’s disease, which involves the same pathway concept and explanatory pattern found in explanations of certain psychiatric disorders.

### 2.3.1 Parkinson’s disease

Parkinson’s disease (PD) is characterized by progressive neurological decline involving symptoms of dementia, resting tremor, rigidity, bradykinesia and postural instability. Research indicates that this disease is caused by both genetic and environmental factors. More specif-

ically, PD can be caused by (1) single gene variants, (2) single environmental influences, and (3) various combinations of these factors (Brady, Siegel, Albers, and Price, 2012). Single gene variants are responsible for less than 10% of all PD cases and are referred to as “monogenic” or “Mendelian” forms of the disease. These variants include mutations in at least five genes (SNCA, LRRK2, PINK1, Parkin, DJ-1), any one of which is sufficient to produce the disease (Shulman, De Jager, and Feany, 2011). These factors are referred to as “causal variants” or “causal genes,” in part because they have a “large effect” on disease occurrence. Single environmental factors can also cause PD, including the drug MPTP,<sup>10</sup> various pesticide agents, and even viral encephalitis (Brady et al., 2012, 861). While PD has these single genetic and environmental causes, most cases of the disease result from a combination of genetic and environmental factors (Brady et al., 2012). These factors are referred to as “susceptibility variants,” “risk variants,” and “modifying factors,” as they are not viewed as individually sufficient to produce the disease, but still influence its likelihood of occurring (Lesage and Brice, 2009; Burbulla and Krüger, 2011). These factors are said to have a “small effect” on the disease outcome and include gene-variants, caffeine, antioxidants, and various toxins. These multi-causal forms of the disease are “clinically indistinguishable” and “pathologically indistinguishable” from monogenic forms (Klein and Schlossmacher, 2006, 137).

The downstream influences of these causal factors are represented as pathways, or causal sequences leading from the specific causal factor(s) to the downstream disease phenotype.

Despite the large number of factors that are causally relevant to PD, evidence suggests that these factors all converge on a common causal pathway that leads to the same Parkinsonian phenotype. Researchers claim that PD is a complex disorder with different causes that all influence a “single, complex, pathophysiological pathway” (Kendler, 2013, 1064) and “operate through a common molecular pathway” (Dauer and Przedborski, 2003) to cause the disease. Although the disorder is causally heterogeneous all cases of the disease are thought to be “related pathophysiologically [and]...caused by a common multifunctional pathway” (Corti, Lesage, and Brice, 2011, 1196). The downstream influences of the early genetic and environmental triggers are represented as pathways or causal sequences that lead

---

<sup>10</sup>MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) is a toxic precursor to MPP+ (1-methyl-4-phenylpyridinium), which causes permanent symptoms of Parkinson’s disease.

to the downstream disease phenotype. The upstream causal factors all disrupt either protein degradation or mitochondrial functioning, both of which converge on a common causal pathway leading to the disease. This common causal pathway includes the (i) cell death of DA neurons, (ii) defective downstream neural connectivity, and ultimately the (iii) disease phenotype, represented in Figure 2. In this manner, all genetic and environmental factors that are causally relevant to PD influence one of two cellular processes, which in turn influence a common set of downstream factors that lead to the disease phenotype.

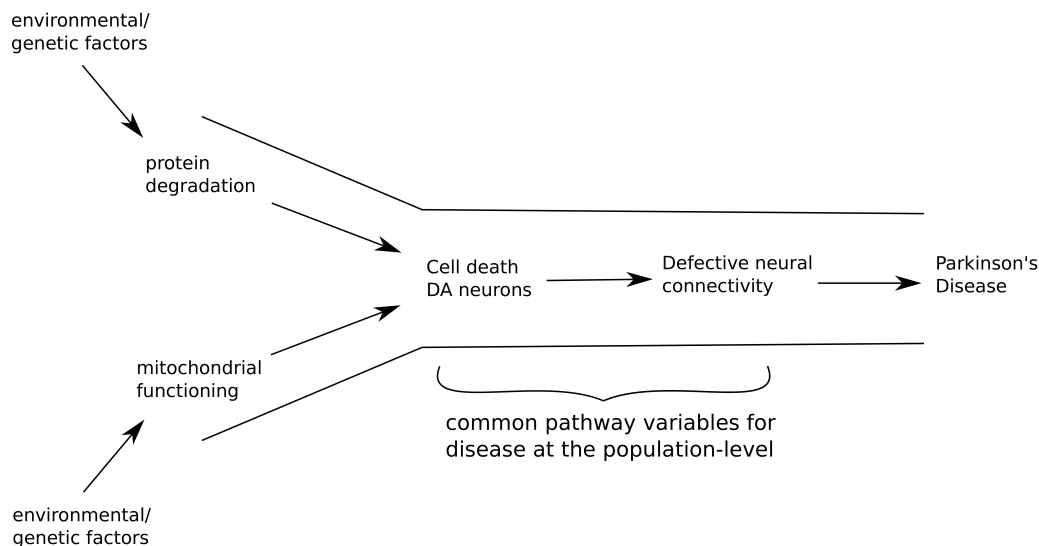


Figure 2: Causal pathways implicated in Parkinson's Disease

This analysis indicates that PD shares a number of important similarities with many psychiatric disorders, particularly those for which pathway explanations have been provided. First, like these psychiatric disorders, the Parkinson's disease phenotype is complex in the sense that it involves multiple symptoms that can manifest in varying degrees. Second, Parkinson's disease involves the same types of causal complexity that are found in these psychiatric conditions. PD is characterized by multicausality as it can be caused by the combined effects of multiple causal factors in individual cases. This is seen in idiopathic cases that involve combinations of genetic and environmental causes. PD is also a causally heterogeneous disease phenotype in the sense that it can be caused by completely distinct causal factors

in different cases of the disease. Different cases of PD can be caused by monogenic factors, single environmental factors, and combinations of genetic and environmental causes working in aggregate. Third, and most importantly, explanations of PD involve the same pathway concept and explanatory pattern as found in the aforementioned pathway-explanations of psychiatric disorders. In the next section, I rely on an interventionist framework to examine how the pathway concept is used to explain disease phenotypes in both situations of causal complexity.

## 2.4 INTERVENTIONIST INTERPRETATION OF PATHWAY EXPLANATION

Woodward’s interventionist account of causation is motivated by the view that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. On this account, to say that  $X$  is a cause of  $Y$  means that, given some background circumstances, an ideal intervention that alters the value of  $X$  causes a change in the value of  $Y$  (Woodward, 2003). To say that this intervention is “ideal” means that it only alters the value of  $X$ , and that this causes a change in the value of  $Y$ , though  $X$  and not through any other route. The causal relationship between these variables can be characterized by a pattern of counter-factual dependance, that captures how specific changes in the value of  $X$  would result in specific values of  $Y$ . Consider a case where the variable  $X$  represents a gene where different values of this variable represent different gene variants {variant 1, variant 2...}. Additionally,  $Y$  is a variable representing a trait where this variable can take different values representing different versions of the trait {trait 1, trait 2...}. To say that  $X$  causes  $Y$  means that changing  $X$  to a specific variant (e.g. variant 1) will cause  $Y$  to take the value of a specific trait (e.g. trait 1). The background conditions are not represented in the  $X$ - $Y$  relationship, but they include required factors for the  $X$ - $Y$  causal relationship to hold (e.g. oxygen levels, temperature, etc.).

The interventionist framework relies on structural equations and directed graphs as tools for representing causal relationships. For example, the causal relationship between  $X$  and

Y can be represented with a directed edge (or arrow) from variable X to variable Y, as seen in Figure 3. In this case, the only variables specified are X and Y and the causal relationship between them is “direct” in the sense that it is not mediated by any variables given the causal system of interest. However, it may be the case that we are interested in more causally complex scenarios where intermediate variables  $\{Z_1, Z_2, \dots, Z_n\}$  span the causal relationship between X and Y, and where X is a “contributing” of Y. A sequence of variables, e.g.  $\{X, Z_1, Z_2, \dots, Z_n, Y\}$ , is a “directed path” or “route” from X to Y, if there is a directed edge from X to  $Z_1$ ,  $Z_1$  to  $Z_2$ ,  $Z_2$  to  $Z_n$ , and  $Z_n$  to Y (Woodward, 2003, 42). As Woodward states, for X to be a contributing cause of Y “there must be a causal chain, each link of which involves a relationships of direct causation, extending from X to Y” (Woodward, 2003, 57). In order to ensure that X is a contributing cause of Y via this causal path and that there are no confounders it must be the case that there is some intervention on X that will change Y when all other variables under consideration, which are not on this path, are fixed at some value (Woodward, 2003, 59). For example, consider the situation in Figure 4 where there is a causal path from X to Y with intermediates  $\{Z_1, Z_2, \dots, Z_n\}$ , there is a causal path from X to Y with intermediate  $\{W\}$ , and there is a directed edge from V to Y. In order for X to be a contributing cause of Y it must be the case that an intervention on X causes a change in Y when all off-path variables—i.e. W and Y—are fixed at some value.<sup>11</sup> This ensures that the intervention on X contributes to Y through the causal path in question.



Figure 3: X as a direct cause of Y

---

<sup>11</sup>For more on this see (Woodward, 2003, 60)

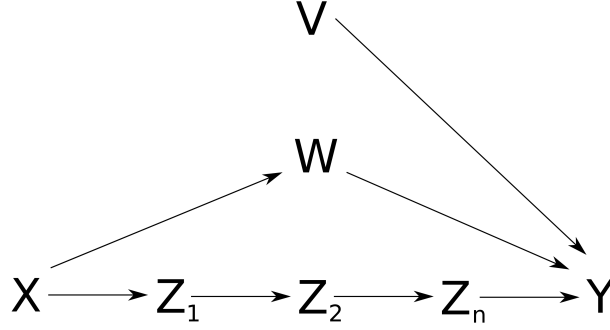


Figure 4: Three causal paths to Y

#### 2.4.1 The pathway concept: two cases of causal complexity

At first glance, there are a number of apparent advantages to interpreting the PD and psychiatric disorder pathway concept with the interventionist account. First, the pathway concept appears to share a number of similarities with the interventionist notion of a “causal path” as represented with the directed graph framework. Furthermore, this account accommodates three features of the pathway concept in the PD case, viz. that these pathways represent a causal sequence, abstract from significant biological detail, and span traditionally accepted biological levels. This account captures how sequences of variables are causally related to each other and the final outcome of interest. These sequences consist of discrete causal steps, which can be viewed as relationships of direct causation. These causal relationships mediate the upstream and downstream variables on these pathways, where upstream variables are contributing causes for a final downstream effect of interest. Furthermore, unlike other theories of causation, the interventionist account can accommodate the common use of pathway variables to represent causal relationships that lack physical connection.<sup>12</sup> This is a common situation in genetics as gene mutations often result in the complete absence of downstream protein products that cause severe downstream effects. An account of causal

<sup>12</sup>More specifically, certain mechanistic theories of explanations require causal relationships have a physical connecting process, which fails to account for cases of absence causation (Machamer and Bogen, 2011) (Waskan, 2011).



explanation that views all causal relationships as characterized by physical connection, is limited in accounting for the fact that scientists in this domain view such gene mutations as causing downstream effects. The interventionist framework also accommodates the manner in which causal pathways abstract from biological details, because it focuses on specific features of causal relationships without representing other biological or non-causal detail. A final advantage of the interventionist account is that it does not restrict causal relationships to single biological levels. This allows it to capture the notion of a pathway in these contexts, where pathways can span biological levels and can include causal variables from different biological levels.

### 2.4.2 Multicausality

Consider the first type of causal complexity, which I call multicausality, where a token instance of some disease is caused by the combined effects of multiple causal factors. Multicausality has been viewed as problematic for explanation in psychiatric genetics for at least two reasons. First, it conflicts with the monocausal model of disease, which maintains that diseases have clear single causal factors. Although monocausality is often viewed as an unrealistic expectation for disease causation, it is still thought to represent the “hard medical model” and the standard of modern medicine (Kendler, 2011). Furthermore, this assumption has been extremely influential in psychiatric genetics. This is indicated by the common pre-GWAS assumption that psychiatric disorders were caused by single gene variants of “large effect” and the profound disappointment of the field when ensuing evidence directly contradicted this (Kendler, 2011; Cannon and Keller, 2006). A second reason for why this type of multi-causality has been viewed as problematic, is that without an understanding that such causal factors are related to each other to produce the phenotype, there is reason to think that they may represent unrelated factors that are merely correlated with the phenotype.

The interventionist account further clarifies these expectations and the use of the pathway concept in cases of multicausality. The monocausal disease model works well for certain cases of PD but not others. What explains why it is appropriate to appeal to single causes in certain cases, but not others? In cases where single factors are cited as the cause of PD, these factors

exhibit a stable relationship to the phenotype in the sense that this relationship holds over a large range of changes in background conditions. If a patient has one of these Mendelian gene variants or environmental exposures, this ensures that they will manifest the disease in situations where background conditions vary significantly. Identifying factors with this type of stability is extremely advantageous, because it provides a specific single variable that is responsible for the phenotype and that can be intervened upon to alter the disease phenotype. This feature of causal relationships is described by Woodward’s concept of “stability” and Kendler’s concept of the “strength,” which both authors indicate is a common feature of paradigmatic causal relationships.<sup>13</sup> This is consistent with the interventionist view that causal relationships are potentially exploitable for manipulation and control, because stable causal relationships allow for such control in a wide range of circumstances. This view is also represented by Murphy’s description of “fundamental explanations,” which he claims appeal to single causes of disease that are “robust” in the sense that they continue to hold with changes in surrounding conditions (Murphy, 2006, 109). Murphy views single-gene diseases as an example of this type of explanation and states that fundamental explanations will not be found for the majority of mental illnesses (Murphy, 2006, 109).

Common monocausal situations not only identify *stable* causal relationships, but obviously they also identify single *specific* causal factors for some effect of interest. Both of these features are valuable for providing causal explanations because they provide clear single locations that explain and allow for control over the phenomena of interest. Attention to stability and specificity can clarify some of the causal reasoning that is operative in psychiatric genetics. When gene variants meet both stability and specificity, like the monogenic and Mendelian gene variants for PD, researchers refer to such factors as “causal variants” (Brady et al., 2012). However, when the factors fail to meet these standards, in cases where there are many gene variants of “small effect,” such factors are not called causal variants, but instead “susceptibility genes,” “risk factors,” or “modifiers.” Thus, the features of stability and specificity appear to identify and characterize the gene-phenotype relationships that researchers in this domain view as causal. This makes sense (from a normative standpoint)

---

<sup>13</sup>For more on Woodward’s discussion of stability see: (Woodward, 2006) and (Woodward, 2010); and for Kendler’s notion of strength see: (Kendler, 2011). Empirical evidence also supports the role of stability in causal reasoning (Lombrozo, 2010).

because such features facilitate the identification of causes that allow for control over some effect of interest.

In situations of multicausality, the pathway concept restores both stability and specificity to situations where these features were originally absent. In integrating the multicausal factors for some token effect of interest, the common pathway provides both a (1) single location that is specific to the effect, and (2) a more stable causal relationship by including input from all relevant causal variables. The common pathway strategy allows for the identification of a single causal pathway in situations that lack a single causal variable. The common pathway provides a single location for researchers to appeal to and target when explaining and treating disease phenotypes. Furthermore, identifying the convergence of multicausal factors on a common process diminishes concerns that such identified factors are unrelated associations to the effect, as evidence suggests their interaction along the pathway. Finally, appealing to the common pathway provides a more stable relationship to the phenotype, because it can capture the causal input of all relevant upstream causes upon which the effect is dependent. As the specific values of the upstream factors can be translated into specific influence on the pathway and values of the pathway variables, this provides a more stable relationship to disease in specifying the required causal influences for single cases of the disease. As all upstream causal factors influence a common pathway, this ensures their causal relation to each other, and allow abstraction from these upstream causal factors by focusing on their downstream influence. This is represented in Figure 5. This can allow for a shift from citing upstream causes to the common pathway in explaining the disease. The interventionist account clarifies why stable and specific causal relationships are valued and how the pathway concept restores these features in situations of multicausality.

### **2.4.3 Causal heterogeneity**

Now I turn to the case of causal heterogeneity where the same disease phenotype is caused by completely distinct causal factors in different patients. This is seen in comparing the various types of monocausal cases of PD where different single genetic and single environmental factors cause different cases of the same disease. Within the psychiatric genetics commu-

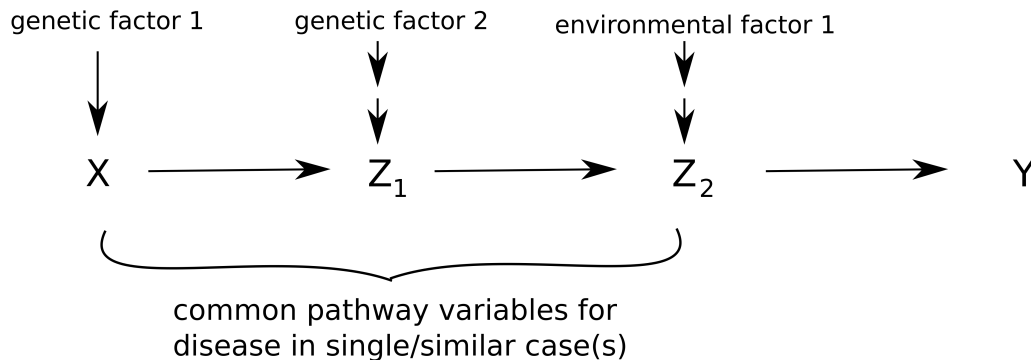


Figure 5: Common pathway in a situation of multicausality

nity, causal heterogeneity has been viewed as much more problematic and challenging for disease explanation than multicausality. Causal heterogeneity conflicts with the strongly held view that human diseases have specific shared causal etiologies, such that all patients with a certain disease have it as a result of some shared causal process. This is such a strongly held view that the identification of a shared causal etiology it is used to justify divisions between disease categories and even “validate” psychiatric conditions as legitimate disorders (Kincaid and Sullivan, 2014). This view provided significant motivation for the common disease-common variant (CD-CV) hypothesis, in pre-GWAS psychiatric genetics, which maintained that common diseases likely shared causal gene variants. Evidence of the causal heterogeneity of psychiatric disorders has been viewed as discrediting this hypothesis and has had a profound impact on psychiatric genetics, where this hypothesis was largely accepted. This evidence has resulted in a “changing landscape” in this domain and a situation where the “earth is shifting beneath psychiatric genetics” as the field comes to terms with the causal heterogeneity that characterizes these disorders (McClellan and King, 2010b, 2523). What is clear, is that the requirement that disease explanation cite some shared causal etiology has been viewed as conflicting with evidence of causal heterogeneity.

Is it correct to view causal heterogeneity as problematic for disease explanation in indicating the lack of shared causal etiology? An answer to this depends on what is meant by shared causal etiology, which is rarely clear in the philosophy of psychiatry (or psychiatric

genetics) literature. The interventionist account provides a natural way to understand the general motivation behind the shared etiology requirement. On this account, explanations of a phenomenon of interest appeal to factors that “make a difference” to the phenomenon. Cases of causal heterogeneity may seem to violate this, because while the explanandum is the shared disease phenotype at the population-level, there is no clear shared difference-maker for all cases of disease in the population, but rather many different heterogeneous difference-makers. Each of these heterogeneous causes can be cited to explain sub-groups of patients with the same disease phenotype, but they cannot be cited to explain all cases of the disease as they do not “make a difference” to all cases of disease. This is the way in which it may make sense to view shared causal etiology as required for explaining a type-level disease phenotypes and how cases of causal heterogeneity may seem to violate such a requirement.

However, while attention to these heterogeneous causes implies a lack of shared causal etiology, in the case of PD heterogeneous causes of the disease are integrated on the basis of their downstream influence on a common pathway. This common pathway is viewed as representing a shared causal etiology for the disease, despite the heterogeneity of upstream causes. Recall that PD can be caused by (1) single gene variants, (2) single environmental factors, or (3) various combinations of these factors, which influence either protein degradation or mitochondrial functioning as shown in Figure 2. Each case of disease can be thought of in terms of an intervention on some upstream genetic or environmental factor. This intervention on an upstream variable influences intermediate variables along a causal pathway leading to the disease phenotype. This causal pathway resembles the causal path notion in the directed graph representation used in the interventionist framework. The variables along the pathway are related by difference-making relations like the direct causal relationships for every link in the chain of a directed path. The upstream cause can be thought of as triggering the flow of difference-making information down the pathway. However, unlike a situation where distinct pathways converge on some effect, the common pathways in the case of PD overlap and share intermediate causal variables. All heterogeneous upstream causes of PD converge on the cell death of DA neurons and defective neural connectivity, which represent a common pathway to PD. Researchers cite this common pathway in explaining PD at the population-level, because it contains causal factors that “make a difference” to

the disease in the entire population. This is further evidenced by the fact that such common pathway variables represent locations that can be intervened upon to treat the population-wide disease. This is valuable from the medical perspective because it affords the potential for treating any case of the disease without knowing anything about the specific upstream causes, that make be unique to an individual patient.

In the case of PD, the common pathway identifies a shared causal etiology despite the causal heterogeneity of upstream genetic and environmental factors. Kendler discusses this approach in the context of psychiatric disorders:

That is, the etiological pathways of numerous basic etiologic processes would pass through a single bottle-neck—perhaps at the level of systems neuroscience or neuropsychology—on their way to causing a psychiatric disorder. We could then use that bottleneck—the place where the diverse etiologic pathways come together—to define our disorders. Note, however, that this approach would not define disorders etiologically in traditional terms. Behind the bottleneck, on which we would base our diagnoses, a range of different etiologic mechanisms would be operative (Kendler, 2011, 17).

The PD case fits this etiological model because it involves a single-bottle neck or common causal pathway that all etiological processes of PD pass through despite their different upstream causes. One of the reasons why pathways can be used to represent “a range of different etiologic mechanisms” is because they abstract from biological details that differ among these processes. They often include minimal information about biological surroundings, intermediate causes, and required factors at each step. For example, the common pathway in PD includes only cell death of DA neurons and defective neural connectivity—their is no information regarding the different causes of DA neuron death, the different molecular ways that this can lead to defective neural connectivity, or the other factors that mediate causal steps in this process. This is consistent with the ordinary use of the term “pathway” to represent some fixed route or road that can be traveled or instantiated by different causal details.<sup>14</sup> Furthermore, the bottleneck or common causal pathway that Kendler discusses can span biological levels, which is also a feature that characterizes the pathway concept in this context.

This analysis suggests that situations of causal heterogeneity do not imply a lack of shared

---

<sup>14</sup>Interestingly, the scientific use of the pathway term is often used with other “road” or “route” type terminology, like flow down a pathway, traffic along a pathway, and information traveling down a pathway.

causal etiology. Some pathway diseases are causally heterogeneous at the lower levels of gene variants, yet homogeneous at the higher levels of cellular functioning and neural circuitry. The identification of common etiology at different levels is suggested by Schaffner who states that “investigators need to be attentive to the possibility of common pathways emerging at any stage (early, intermediate, and final) in the temporal evolution of a reticulate network and involving multiple causes and complex “crosstalk” (Schaffner, 1998, 241). In these cases, the level of the common pathway dictates the level of detail cited in explanations of type level disease phenotypes.

#### **2.4.4 Causal complexity and challenges for explanation**

Evidence continues to suggest that an extreme degree of causal complexity underlies many psychiatric disorders. Two common forms of this complexity include multicausality and causal heterogeneity. These types of causal complexity are commonly viewed as evidence against the possibility of explaining psychiatric disorders. Such claims often target causal heterogeneity and invoke the (1) lack of some shared causal etiology or (2) the multiple realization of the higher level disease phenotype by differing lower level details. My analysis of the common pathway strategy indicates that these claims are incorrect. With regard to the common etiology objection, the common pathway strategy indicates that situations of causal heterogeneity do not prevent causal explanation or indicate the lack of some important sense of shared causal etiology. In these situations, a shared causal etiology is identified, although it is downstream or at a higher level than the heterogeneous genetic factors. These complications and the frequency with which this objection is found in the philosophy of psychiatry literature suggest that there is a serious need to clarify what is meant by “shared causal etiology” and what type of requirement or role it serves in disease explanation. In this literature there is an assumption that the identification of causal etiology will reveal “true” divisions between disorders. My analysis suggests that such claims involve an unrealistic understanding of causal etiology, which fails to appreciate its dependence on the explanatory target of interest. With regard to the multiple realizability objection, the common pathway strategy indicates that researchers to not view the multiple realizability of such phenotypes

as preventing explanation or as preventing appeal to lower level details. Multiple realizability can help clarify why lower level heterogeneous causes are not cited in such explanations, but it is silent on whether downstream shared etiology is present or not, and if so, at which level it will emerge. Furthermore, researchers in psychiatric genetics are very comfortable discussing causal relationships that span the mind brain divide. When they can explain psychiatric disorders they often do so by appealing to lower level neurological details of the brain.

## 2.5 PATHWAYS AND MECHANISMS

In the philosophical literature, it is often assumed that psychiatric disorders are defined on the basis of shared causal etiologies ([Murphy, 2014](#); [Caspi and Moffitt, 2006](#); [Zachar, 2014](#)). Some philosophers of psychiatry have claimed that this notion of shared causal process is captured by the philosophical notion of a mechanism and associated accounts of mechanistic explanation ([Kendler, Zachar, and Craver, 2010](#); [Craver, 2009b](#)). Philosophers might assume that the pathway concept should be interpreted mechanistically and others have explicitly claimed that the notion of a “pathway” is clearly accommodated by a mechanistic framework ([Craver, 2009a](#), 3). Mechanist theories claim that explanations in biomedicine appeal to the mechanisms that underlie the scientific phenomenon of interest. While there are different versions of these theories, most of them define mechanisms as the underlying component parts of a system and the features, activities, and organization of these components that are relevant to the production of a particular phenomena of interest ([Machamer, Darden, and Craver, 2000](#)). Mechanism components are said to be related in a constitutive, or part-whole manner, where factors in the explanans constitute (and cause) the explanandum. The explanandum fixes some higher-level phenomena that is explained by referring to causally related lower-level mechanism components that produce it. This leads to a situation where lower-level mechanisms are nested within higher-level mechanisms to create “levels of mechanisms,” which proponents of this view claim “captures the central explanatory sense in which explanations in neuroscience (and elsewhere in the special sciences) span multiple



levels” (Craver, 2009a, 163). This notion of “levels of mechanisms” is represented in Figure 6 (Craver, 2009a, 189). This sentiment is reiterated by Bechtel and Craver who claim that purported lower-level causes of higher-level phenomena (or ‘bottom-up causation’) are best understood as mechanistically mediated effects “where the constitutive relations are inter-level, and the causal relations are exclusively intralevel” (Craver and Bechtel, 2007, 547). Thus, mechanisms are set of causally related components at a particular biological level, where inter-level relations are constitutive relations between mechanisms at different levels.

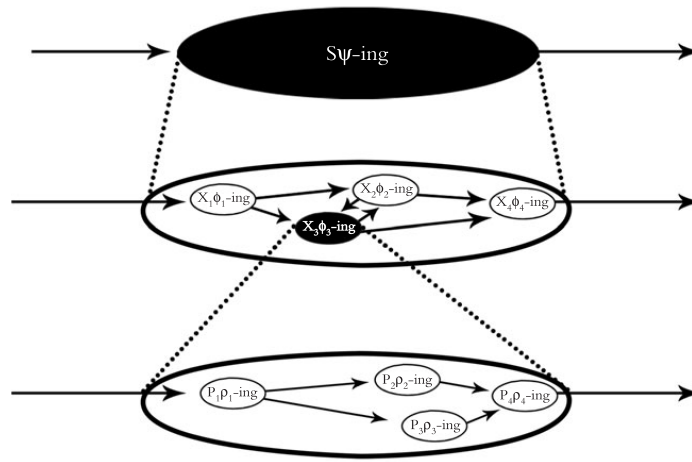


Figure 6: Craver’s representation of levels of mechanisms and constitutive relations between levels (Craver 2007, p. 189)

This mechanist account cannot accommodate the pathway concept analyzed in this chapter, because this pathway concept is not characterized by causal relationships that are constrained to particular “levels.” Alternatively, the pathway concept refers to causal relationships that span biological levels and abstract from significant “lower-level” or mechanistic detail. For example, the common causal pathway in PD spans the biological levels of cellular death and neural network rearrangement (i.e. it involves the (1) cell death of DA neurons and (2) defective downstream neural connectivity). This pathway is cited in explaining the disease and it cannot clearly be accommodated by a theory of causal explanation where causal relations are restricted to biological levels. Furthermore, the abstract nature of these

causal pathways is a useful feature in the contexts in which they are used, because it allows them to represent a shared causal route that is traveled by many mechanistically distinct instantiations. This is supported by how scientists distinguish between these two terms. In cases where scientists claim to have identified biological pathways, they further clarify whether they know of the mechanisms that underlie or make-up the pathway or not. Oftentimes, they state to have identified a biological pathway without knowing the mechanisms that make it up, although they view the pathway as providing a helpful step towards elucidating them. Additionally, they state that many biological pathways can have different mechanistic details, e.g. different enzymes or co-factors that operate at the same causal step or different intermediate substrates in different situations. If scientists distinguish between these concepts and if the philosophical concept of a mechanist poorly accommodates the scientific use of the pathway concept, these provide reasons for further elucidating the differences between these causal concepts and why they may be used in certain situations and not others.

This analysis suggests that the pathway concept used in these cases is not accommodated by the mechanist account of explanation, and that scientists likely use the terms “pathway” and “mechanism” to refer to causal structures that differ in important ways. As the most successful mechanist theories rely on an interventionist account of causation, this analysis suggests a promising and different way of studying and understanding the complex causal concepts appealed to in biological explanations: complex causal concepts, e.g. mechanism, pathway, etc., might refer to different causal structures that can be understood as being made-up of the same “building-block” notion of causation. These different causal structures may be more useful, or appropriate in various way, for capturing causal relations in particular biological contexts. Further examining the differences between complex causal concepts, and why they are invoked in certain situations and not others can help to clarify important aspects of biological explanation. This is important for our understanding and explanation of psychiatric disorders, because attempts to specify the requirements of psychiatric classification (or disorder validation) claim that what is required is the identification of some shared or specific causal process (Stein, 2014, 68).

## 2.6 CONCLUSION

In this chapter I have examined an explanatory pattern in psychiatric genetics where the notion of a pathway is used to explain disorders in situations of causal complexity. I have indicated how the pathway concept is used to explain disease phenotypes that involve multicausality and causal heterogeneity. In these cases the identification of a common pathway allows for the integration of factors that are all causally related to some shared phenomenon of interest. Pathway explanations reveal that there are different types of shared causal etiology: a disorder can be genetically heterogeneous, but still involve etiologic homogeneity across molecular, cellular, neural network, and other levels. The type of shared causal etiology preferred will depend on multiple considerations including the explanatory target of interest and the type of causal complexity encountered.

### 3.0 DYNAMICAL MODELS AND EXPLANATION IN NEUROSCIENCE

#### 3.1 INTRODUCTION

This section contains further description of Kaplan and Craver’s claims regarding explanations in neuroscience, including their 3M constraint and the claim that the explanatory power of a model increases as it includes more mechanistic detail. They direct these claims at mathematical models in neuroscience and use them to distinguish between explanatory models and those that merely provide descriptions or predictions.

According to Kaplan and Craver, all explanations in neuroscience appeal to mechanisms as models in this field “carry explanatory force to the extent, and only to the extent, that they reveal (however dimly) aspects of the causal structure of a mechanism” (Kaplan and Craver, 2011, 602). They define mechanisms as the underlying component parts of a system and the features, activities, and organization of these components that are relevant to the production of a particular phenomena of interest (Kaplan and Craver, 2011, 605). Explaining this phenomenon requires citing all and only those actual components and activities that underlie and produce it. For example, an adequate explanation of neural firing (or the action potential) appeals to the relevant biological entities and activities that underlie and produce this firing. These biological entities include the relevant ion channels, ions, and the  $Na^+/K^+$  pump, while the activities describe what these entities do, e.g. their attraction, blocking, diffusion, etc. (Craver, 2008, 1025). As an account of causal explanation, the mechanist position depends on the rationale that explaining a phenomena of interest requires citing the causal factors that produce it. In other words, it requires that the explanans invoke factors that are causally relevant to the explanandum. If a model merely describes or predicts the explanandum, without citing the causal factors that produce it, the model is regarded as

non-explanatory.<sup>1</sup>

There are two central claims that Kaplan and Craver make regarding the explanatory status of mathematical models in neuroscience. The first is their model-to-mechanism-mapping (3M) constraint, which states:

(3M) In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism (Kaplan and Craver, 2011, 611).

Although this statement of 3M explicitly addresses models in cognitive and systems neuroscience, Kaplan and Craver extend it to all models in neuroscience.<sup>2</sup> Their 3M constraint specifies two mapping-relations that must be met between the model and a target system in order for the model to be explanatory. The first maps the variables of a model to components within the system and the second maps dependencies among variables in the model to causal relations among components in the system. These criteria are intended to ensure that the model accurately represents the “internal aspects of the system” (Kaplan and Craver, 2011, 616). However, the degree to which a model needs to fulfill 3M in order to be explanatory is not made entirely explicit in their work. They indicate that models need not completely map to the target system or refrain from idealizations and abstractions to be explanatory. Kaplan states that “3M requires only that *some* (at least one) of the variables in the model correspond to at least *some* (at least one) identifiable component parts and causal dependencies among components in the mechanism responsible for producing the target phenomenon”

---

<sup>1</sup>The distinction between explanatory models and those that are merely descriptive or predictive has received significant attention in the philosophical literature on explanation, particularly in the context of the biological sciences. Woodward thoroughly addresses this distinction in his account of causation, which Kaplan and Craver rely on in specifying their notion of a causal mechanism (Woodward 2003; Kaplan and Craver 2011, 602, 613). Woodward’s interventionist theory of causation maintains that explaining a phenomenon involves citing the causes that “make a difference” to the phenomenon, in the sense that if these causal factors were changed they would change the explanandum in various ways. Descriptive and predictive models are not explanatory because they do not cite factors that are causally relevant to the explanandum, but merely re-describe or predict the explanandum, respectively.

<sup>2</sup>They focus on cognitive and systems neuroscience to argue that mechanistic explanation is the unique form of explanation in higher-level neuroscience, which they take to have already been established for lower-level neuroscience (Kaplan and Craver, 2011, 602-3). In a separate paper, Kaplan argues for the 3M criteria in the context of computational neuroscience (Kaplan, 2011). For a helpful discussion of explanation in computational neuroscience and response to Kaplan’s paper, see (Chirimuuta, 2013).

(Kaplan, 2011, 347-8; emphasis original). In this manner, the 3M constraint is stated such that it requires only a minimal amount of mapping from the model to the target system.

The second main claim that Kaplan and Craver make is that among models meeting 3M, the explanatory power of a model increases as it includes more relevant mechanistic detail (Kaplan, 2011, 347). According to Kaplan:

As one incorporates more mechanistically relevant details into the model, for example, by including additional variables to represent additional mechanism components, by changing the relationships between variables to better reflect the causal dependencies among components, or by further adjusting the model parameters to fit more closely what is going on in the target mechanism, one correspondingly improves the quality of the explanation (Kaplan, 2011, 347).

As including increasing amounts of detail into a model further reveals the causal structure of the mechanism, it increases the explanatory status of the model. Kaplan and Craver sometimes refer to this claim as a “fact” and at other times a “highly plausible assumption” (Kaplan and Craver 2011, 613; Kaplan 2011, 347). In either case, it is presented as a complement to their 3M constraint. This more-details claim provides a natural way of assessing the degree to which a model meets 3M or maps onto a causal mechanism. A more detailed mechanistic model, with a higher degree of mapping, will provide a better explanation because it will be able to answer a wider range of questions about the physical system of interest.

Kaplan and Craver defend a strong mechanist position in an ongoing debate about the explanatory status of dynamical models in neuroscience. They use their position to argue against the claim that dynamical models can be explanatory when they do not reveal the causal structure underlying system-level dynamics. However, as dynamical models often contain variables that represent macroscopic and behavioral features of neural systems, these variables do not always appear to map onto mechanisms in the 3M sense. In these cases, Kaplan and Craver claim that these variables “are not components in the sense of being the underlying parts of the mechanism” and merely provide mathematically compact characterizations of system-level behavior (Kaplan and Craver, 2011, pp. 615-614). They state that these dynamical models provide at best *descriptions* or *predictions* of the behavior of complex mechanisms and that those who consider them explanatory “fundamentally misidentify

the source of explanatory power in their models” (Kaplan and Craver, 2011, 602). This criticism is directed towards those who have argued for distinctly dynamical, non-mechanistic explanations in neuroscience, which has been argued, most notably, by Stepp, Chemero, and Silberstein.<sup>3</sup> Unfortunately, these dynamicist arguments have remained susceptible to such an objection, because they have continued to reference the predictive success of these models without providing another clear sense in which they are explanatory. The strong mechanist position has likely benefitted from the fact that these arguments for non-mechanistic dynamical explanation have not been viewed as entirely successful.

With this description of Kaplan and Craver’s mechanist position, I move on to providing some background on dynamical systems neuroscience. In this section, I first discuss how neural excitability is understood and modeled with the dynamical systems approach. To do this I characterize neural excitability from a molecular perspective and contrast this with the dynamical systems perspective. After clarifying certain aims of dynamical modeling I provide an example of an explanatory dynamical model in neuroscience. I indicate why this dynamical model is explanatory and what led neuroscientists to seek the explanation it provides.

### 3.1.1 Dynamical models in neuroscience

A major topic of study in neuroscience is the excitability of neurons as this is important for understanding how they transmit information. From the molecular perspective neural firing, or the action potential, is explained with a generic neuron model consisting of voltage-gated ion channels sensitive to  $Na^+$  and  $K^+$ . When a neuron receives a strong enough signal a number of things happen in succession that cause it to fire. First, the sodium channels

---

<sup>3</sup>For example, Silberstein and Chemero claim that dynamical models allow for the prediction of qualitative behavior and that “[i]f models are accurate enough to describe observed phenomena and to predict what would have happened had circumstances been different, they are sufficient as explanations” (Silberstein and Chemero, 2008, 12). Stepp, Chemero, and Turvey argue that dynamical models are genuinely explanatory and claim that, similar to covering-law explanations, “dynamical explanations show that particular phenomena could have been predicted, given local conditions and some law-like general principles” (Stepp, Chemero, and Turvey, 2011, 432). In a recent paper, Silberstein and Chemero provide a different argument for non-mechanistic explanation in neuroscience by claiming that some explanations fail to meet the mechanistic requirements of localization and decomposition (Silberstein and Chemero, 2013). In footnote 17, I discuss their position in more detail.

open quickly and  $Na^+$  rushes into the cell causing the membrane potential to increase. This results in depolarization of the neuron and the upstroke of the action potential. Shortly after this depolarization, the potassium channels open and  $K^+$  rushes out of the cell, while the sodium channels begin to close, decreasing the influx of  $Na^+$ . These events cause the membrane potential to decrease, which contributes to the repolarization of the neuron and downstroke of the action potential. The action potential travels down the length of the neuron and constitutes a single firing event.

In dynamical systems neuroscience, neural excitability is understood and modeled in a different way: the main aim is to study qualitative features of neural systems irrespective of their fine-grain molecular details. Qualitative features of neural systems are studied by analyzing the graphical and topological structures of dynamical models that represent these systems. A dynamical model is a mathematical model that describes how variables representing a particular system evolve with time. In neuroscience it is common to model neural excitability in this way with coupled differential equations. For example, consider the following two-variable dynamical model:

$$\dot{V} = f(V, n) \tag{3.1}$$

$$\dot{n} = g(V, n) \tag{3.2}$$

This is a system of coupled differential equations that describe how  $V$  and  $n$  change over time. Here,  $V$  is the excitation variable, which represents neural factors responsible for depolarization, and  $n$  is the recovery variable, which represents neural factors responsible for repolarization. The functions  $f$  and  $g$  describe the evolution of the two-dimensional state variable  $(V(t), n(t))$ . With this two-variable model the dynamical system can be represented graphically, as shown on the phase plane in Figure 7. In this figure  $V$  is plotted along the x-axis and  $n$  is plotted along the y-axis. To each point  $(V, n) \in \mathbb{R}^2$  there is a corresponding vector whose x component is  $\dot{V}$  and whose y component is  $\dot{n}$ . The vector field plotted shows  $(\dot{V}, \dot{n})$  at each  $(V, n)$ .

Graphical analysis of the vector field on the phase plane can provide information about the system that may not be obvious from the differential equations alone. For example, a



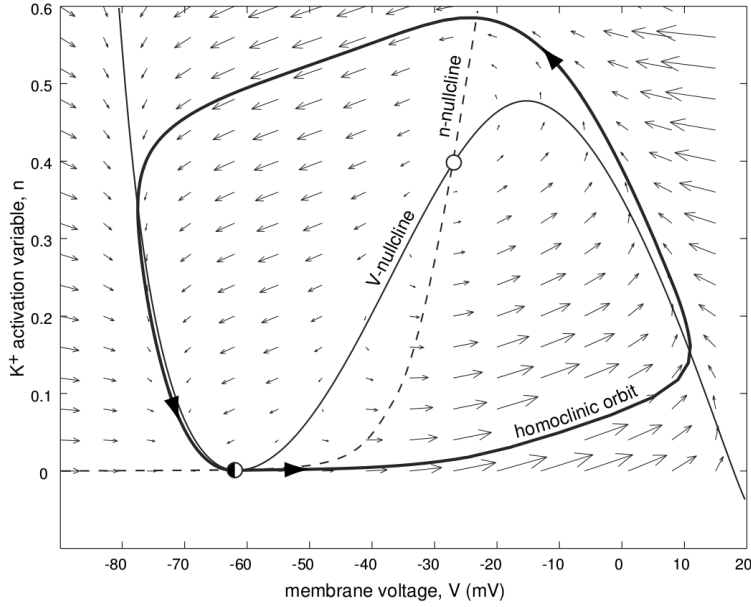


Figure 7: Phase plane with vector field (Izhikevich (2007), p. 113)

solution to the system of equations can be obtained from an analysis of the figure, as it is the curve  $(V(t), n(t))$  on the phase plane tangent to the vector field. The significance of a solution to the system of equations is that it gives a full picture of how  $V$  and  $n$  change over time. This solution and its portrayal as a trajectory corresponds to a characterization of neural firing. Counter-clockwise movement on the trajectory tracks changes in  $V$  and  $n$  throughout the action potential and the completion of this trajectory represents a single firing of the neuron.<sup>4</sup>

In dynamical systems neuroscience the qualitative features of neural systems are often studied without reference to their fine-grained molecular detail. There are two main reasons for this. First, as graphical representations and qualitative features are exhibited by systems of differing molecular details, explaining these qualitative features does not depend on a shared physical structure. For example, the phase plane in Figure 7 represents the qualitative behavior of neurons that differ in their physical structure, i.e. in terms of their ion

<sup>4</sup>For more on graphical representations of neural excitability see (Ermentrout and Terman, 2010; Izhikevich, 2007).

channels, ion permeabilities, etc. The fact that physically distinct neural systems can exhibit the same qualitative behavior motivates the view that this behavior is, in a sense, independent of any specific molecular microstructure. Hoppensteadt and Izhikevich express this sentiment when they state that “[b]ehavior can be quantitatively different, but qualitatively the same” (Hoppensteadt and Izhikevich, 1997, 33).<sup>5</sup> As the focus in dynamical systems neuroscience is on studying and explaining the qualitative behavior of neural systems, the physical differences among systems that exhibit these behaviors are rarely referenced (and sometimes the full extent of these differences are unknown). A second reason for this inattention to molecular detail is that preferred graphical analyses, which concisely represent the comprehensive behavior of neural systems, constrain the number of variables that can be implemented to characterize these systems. This requires the use of simple models that abstract from the molecular details of neural systems, while preserving their system-level behavior. The use of such techniques by Fitzhugh and Nagumo et al. in the early 1960s essentially marks the beginning of dynamical systems neuroscience (Fitzhugh, 1960; Nagumo, Arimoto, and Yoshizawa, 1962). Fitzhugh pioneered this work by reducing the number of variables in the Hodgkin-Huxley model of the action potential so that the system could be “easily visualized” in a phase-space, leading “to a better understanding of the complete system than can be obtained by considering all the variables at once” (Fitzhugh, 1960, 873). He reduced the number of variables in these neural models by exploiting their different time scales and functional effects.<sup>6</sup> This early work explicitly distinguished the qualitative features of neurons and the topological properties of their phase space, from an analysis of their physical constitution. As I discuss in the following subsections, these techniques and the general aims of dynamical systems neuroscience are central to understanding how some models in this field are used to provide explanations.

---

<sup>5</sup>In other words, quantitative differences between neural models indicate physical differences between the systems they represent, even though such models can exhibit the same qualitative behavior.

<sup>6</sup>For Fitzhugh’s use of these reduction techniques see (Fitzhugh, 1960, 1961) and for further discussion of them see (Abbott, 1994; Doi and Kumagai, 2001, 69).

### 3.1.2 Explanatory Dynamical Model: the Canonical Model

In this subsection, I give an example of a dynamical model in neuroscience and present an account of its role in a particular explanation. In this example the dynamical model, referred to as a canonical model, represents the shared qualitative features of a number of physically distinct neural systems. I indicate how this dynamical model is used to provide explanations after discussing the research findings that led neuroscientists to seek these explanations.

In 1948 Hodgkin published important results from his voltage clamp studies of single crab neurons (Hodgkin, 1948). In these experiments he measured the electrical responses of neurons after injecting them with various levels of current. He identified three different types of neural excitability, which he referred to as class I, class II, and class III excitability, a categorization still used today.<sup>7</sup> Class I neurons exhibit a low frequency of firing to low levels of current and smoothly increase their firing with increases in current. Class II neurons begin firing when the current stimuli reaches a higher level and their firing frequency increasing minimally with increases in current, as represented by the step function in Figure 8. The relationship between current introduced into class I and class II neurons and the frequency of their firing response is represented in the frequency-current (F-I) graph in Figure 8. Class III neurons fail to maintain firing in response to current stimuli (and are not depicted in the figure). The qualitative distinctions between these classes is that for class I neurons the frequency-current relationship starts at zero and increases continuously, for class II neurons it is discontinuous, and for class III neurons it is not defined.

These excitability classes identify qualitative features that are shared among large groups of physically distinct neurons. Hodgkin was particularly interested in class I excitability because it had been identified in neurons from many different animals (Hodgkin, 1948, 167). Since his work, neuroscientists have identified class I excitability in many other neural systems, including rat hippocampal neurons, rat cortical neurons, crustacean motor neurons, and the majority of neurons in the mammalian cortex (Tateno, 2004; Jia, Gu, and Li, 2011; Connor, 1975; Cauli, Audinat, Lambolez, Angulo, Ropert, Tsuzuki, Hestrin, and Rossier, 1997). As neurons with class I excitability are found in animals of different biological phyla

---

<sup>7</sup>These categories are sometimes referred to as type I, type II, and type III neuronal excitability (Hoppensteadt and Izhikevich, 1997, 84).

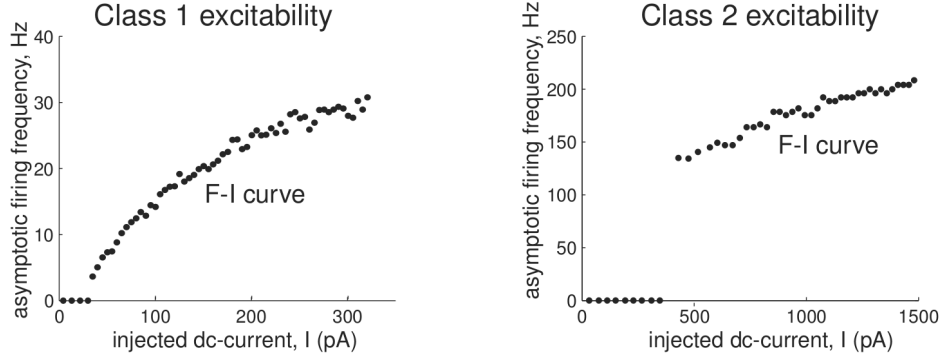


Figure 8: Graph of the frequency-current (F-I) relationship of class I and class II neurons (Izhikevich (2007), p. 14)

and even throughout the nervous systems of single species, it is unsurprising that this class encompasses neurons that differ in their microstructural details. What has been surprising, however, is the astonishing degree of this variation and the complexity of neural structures that has been revealed by recent advances in patch-clamp recording, heterologous expression of cloned channels, and genomic analysis (Bean, 2007). For example, consider mammalian pyramidal neurons, the majority of which exhibit class I excitability. These neurons have three main types of voltage-gated ion channels responsible for excitability, including those selective for  $Na^+$ ,  $K^+$ , and  $Ca^{2+}$ . Of those channels that transmit distinct ions each have an enormous variety of subtypes, for example, there are over 100 molecularly distinct  $K^+$  channels (Vacher, Mohapatra, and Trimmer, 2008). From this large selection of channels a single neuron typically expresses over a dozen different types, which vary in density along the neural membrane and result in many distinct voltage-dependent conductances. These voltage-dependent conductances contribute to the excitability of these neurons and can be comprised of 2-5 different currents each of ion ( $Na^+$ ,  $K^+$ , and  $Ca^{2+}$ ) (Bean, 2007). This indicates a large degree of molecular difference among mammalian pyramidal neurons with class I excitability. The differences between all neurons that share this behavior is, of course, much greater.

Neuroscientists have sought an explanation for why neurons that differ so drastically

in their microstructural details all exhibit the same type of excitability. In this case the explanandum is a behavior displayed by a group of physically distinct systems as opposed to a behavior produced by a single physically unique system. In 1986 Ermentrout and Kopell provided the crucial component of this explanation with their derivation of a canonical model for class I excitability.<sup>8</sup> Their work involves using mathematical abstraction techniques to reduce models of molecularly diverse neural systems to a single model, referred to as a canonical model. The canonical model and abstraction techniques used in this approach explain why molecularly diverse neural systems all exhibit the same qualitative behavior and why this behavior is captured in the canonical model. The explanation for this shared behavior is that when mathematical abstraction techniques are used to abstract away from details of mathematical representations of neural systems, all representations converge onto the same canonical model. In the next subsection, I further describe the abstraction steps, canonical model, and the explanations they provide.

**3.1.2.1 Reducing models of neural excitability** The first step in this canonical model approach involves reducing the number of variables in models of neural excitability. Generally, variables characterizing the dynamics of neural systems are classified into four groups depending on their time scale and effect on membrane potential. These variables include: (1) the membrane potential variable, (2) excitation variables, (3) recovery variables, and (4) adaptation variables (Izhikevich, 2007, 8). Excitation variables include neural factors that contribute to the upstroke of the action potential and firing of the neuron, while recovery variables represent neural factors that contribute to the downstroke of the action potential and recovery of the neuron. Adaptation variables stand for neural features that increase during continued spiking and can alter long-term neural excitability. This classification allows the factors characterizing neural excitability to be collapsed into one of the four variables that together characterize the dominant behaviors of the system.

Models for class I excitability do not contain variables of the fourth type, so our analysis begins with dynamical models characterized by three variables: the membrane potential

---

<sup>8</sup>This model is also called the “Ermentrout-Kopell model” and sometimes the “theta model” (Izhikevich, 2004; Ermentrout, Rubin, and Osan, 2002; Börger, Epstein, and Kopell, 2008).

variable, excitation variable, and recovery variable. A model with these three variables can be reduced to a two-variable model by exploiting differences in the rate of the kinetics between the excitation and recovery variables.<sup>9</sup> As the kinetics of the excitation variable are often much faster than the kinetics of the recovery variable, an idealization is introduced into the model by replacing the excitation variable with the value it quickly approaches (Rinzel and Ermentrout, 1989). This reduces the model to two variables that characterize the macro-level behavior and dynamics of the system: the “new” excitation variable  $V$ , which was formerly the membrane potential variable,<sup>10</sup> and the recovery variable  $n$ . This two-variable dynamical model takes the same form as the coupled differential equations (3.1) and (3.2).

When models of neural excitability are reduced to two variables and represented graphically, those systems with class I excitability all exhibit the same change in topological structure as they transition from resting to sustained firing. This qualitative feature is captured in dynamical systems theory by the presence of a particular kind of bifurcation. In the case of neurons with class I excitability, all exhibit the saddle-node on invariant circle bifurcation (Izhikevich, 2007, 164). This reduction of mathematical models of neural excitability to two-variable models is the first step in the canonical model approach and begins to reveal the shared qualitative features in their topology.<sup>11</sup>

**3.1.2.2 Ermentrout-Kopell Theorem** Identifying this particular bifurcation in all models of class I systems is significant because Ermentrout and Kopell’s theorem for class I excitability proves that all models which exhibit this bifurcation transform into the same

---

<sup>9</sup>The use of scale differences to reduce variables in mathematical models is a well-known approach. For more on this approach, see: (Fowler, 2007; Batterman, 2000).

<sup>10</sup>Once this reduction is performed it is common to refer to the variable for the membrane potential as the “excitation variable.” This is because the membrane potential variable tracks changes in the neural membrane due to current stimuli, which can result in excitation of the neural system.

<sup>11</sup>This first step allows for the representation of system-level behavior in a two-dimensional phase space and serves many important roles in understanding this behavior, e.g. in identifying the particular bifurcation that characterizes the system. However, for the purpose of reducing any model of neural excitability to the canonical model, so long as the system exhibits the saddle-node on invariant circle bifurcation, technically the Ermentrout-Kopell theorem is all that is required (Hoppensteadt and Izhikevich, 1997).

model when they are reduced. They prove this by providing a continuous piecewise transformation, represented by  $h$  in Figure 9, that transforms any one-variable model, among a family of models with this bifurcation, into a single canonical model.

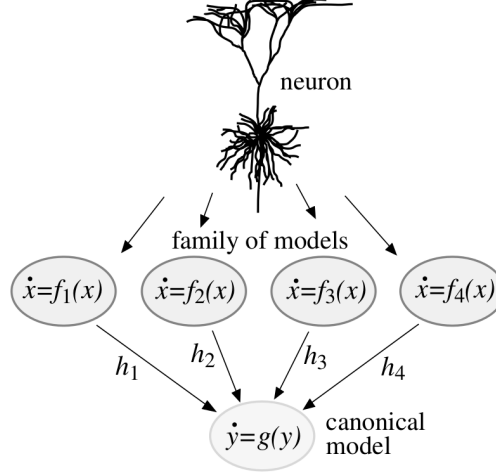


Figure 9: Modeling techniques in neuroscience (Izhikevich (2006), p. 279)

In other words, Ermentrout and Kopell prove that all dynamical systems with the saddle-node on invariant circle bifurcation of the form:

$$\dot{x} = f(x), \quad x \in \mathbb{S}^1, \quad (3.3)$$

can be mathematically transformed into the following canonical model:<sup>12</sup>

$$\theta' = (1 - \cos\theta) + (1 + \cos\theta)r, \quad \theta \in \mathbb{S}^1, \quad (3.4)$$

where  $\theta$  represents the activity of a neural system given a particular current input represented

---

<sup>12</sup>Ermentrout and Kopell's theorem pertains not just to single neurons but also to neural networks. The equations that pertain to neural networks contain an extra term that accounts for the connectivity and interactions between neurons. For these equations see (Hoppensteadt and Izhikevich, 1997, 225). I have chosen the single neuron case for simplicity of presentation.

by  $r$ .<sup>13</sup> Given a particular fixed value of the bifurcation parameter  $r$ , the model describes how the activity of the neural system, represented by  $\theta$ , changes over time by specifying the location of  $\theta$  on the unit circle  $\mathbb{S}^1$ . This is represented in Figure 10, where the location of  $\theta$  on the unit circle indicates whether the neural system is in the rest, threshold potential, spike, or refractory phase. Every completion of the unit circle by  $\theta$  represents a single firing event of the neural system. The model indicates that with small values of  $r$  the neural system remains at rest, represented by the variable  $\theta$  at the rest potential position. Larger values of  $r$  result in continuous firing of the neural system, represented by the continuous movement of  $\theta$  around the unit circle.

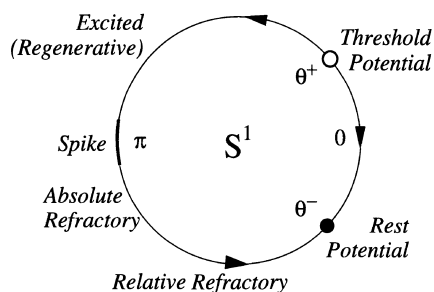


Figure 10: Physiological state diagram of a Class I neural system (Hoppensteadt & Izhikevich (1997), p. 228)

<sup>13</sup>Equations (3.3) and (3.4) use different notational conventions for derivatives. Dots denote derivatives with respect to time, while primes denote derivatives with respect to some other specified variable. In equation (3.4) the prime represents a derivative taken with respect to the variable  $\tau$ , where  $\tau = \sqrt{\varepsilon}t$  and  $\varepsilon$  is a small parameter (Hoppensteadt and Izhikevich, 1997, 225).



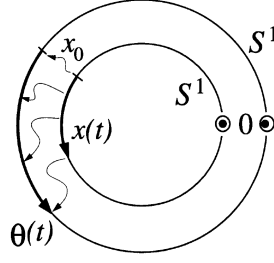


Figure 11: The solution  $x(t)$  of (3.3) is mapped to the solution  $\theta(t)$  of (3.4), the canonical model (Hoppensteadt & Izhikevich (1997), p. 119)

Ermentrout and Kopell’s theorem for class I excitability provides a continuous transformation  $h : \mathbb{S}^1 \rightarrow \mathbb{S}^1$  that converts solutions of (3.3) to solutions of (3.4), represented in Figure 11. This figure shows how any point on the unit circle  $\mathbb{S}^1$  of the family model is represented on the unit circle  $\mathbb{S}^1$  of the canonical model. This transformation preserves the behavior of the original system and ensures that no artifacts or behavior not present in the original system are inherited by the canonical model. Neuroscientists describe this transformation as “extracting some particularly useful dynamical features” from these models, which are represented in the canonical model for class I excitability (Hoppensteadt and Izhikevich, 1997, 115). This approach reveals how all models of systems with class I excitability are transformed into the same canonical model when they are reduced with principled mathematical techniques. This reveals how mathematical representations of class I systems are stable under certain perturbations by abstracting away from details of each model.

One of the more impressive features of this canonical model is that it provides the frequency with which any class I neuron will oscillate given a particular fixed value of  $r$  (Hoppensteadt and Izhikevich, 1997, 227-8). The canonical model approach is valued by mathematical neuroscientists because it provides a rigorous method for gaining information about classes of neural systems which share a particular behavior without obscuring this similarity behind the details of any one system (Izhikevich, 2007, 278). As Izhikevich notes, the “advantage of this approach is that we can study universal neurocomputational properties

that are shared by all members of the family because all such members can be put into the canonical form” (Izhikevich, 2007, 278). Furthermore, as this canonical model approach pertains not just to single neurons, but also to neural networks, it indicates the relevance of this explanatory approach to both cellular and systems-level neuroscience.<sup>14</sup>

It is worth emphasizing that this approach depends crucially on both the canonical model and mathematical abstraction techniques that relate it to models of distinct neural systems. Referring to the canonical model alone could be viewed as merely describing or predicting the behavior of class I neurons, as opposed to explaining it. The canonical model approach, however, including the canonical model and abstraction techniques, does more than just describe or predict the excitability of class I neurons. It explains *why* physically distinct neural systems all share the same behavior by showing that principled mathematical abstraction techniques—which preserve qualitative behavior—can be used to reduce all models of these distinct systems to the same canonical model. These abstraction techniques involve exploiting time scale differences to introduce idealizations into models and transforming systems into simpler models that are topologically equivalent. This approach provides an explanation for this shared behavior – when principled mathematical techniques abstract from the details of different systems, they can all be simplified into the same canonical model that exhibits this behavior.

### 3.2 ANALYSIS OF THE CANONICAL MODEL APPROACH

In this section, I examine whether Kaplan and Craver’s 3M constraint and claims regarding detailed models—which they created to account for explanatory mathematical models in neuroscience—can accommodate the explanations provided by the canonical model approach. I argue that their mechanist criteria and framework cannot account for this type of explanation. I then describe Batterman’s account of minimal model explanations and argue that it well characterizes the explanatory structure of the canonical model explanation by eluci-

---

<sup>14</sup>For more on Hoppensteadt and Izhikevich’s discussion of the canonical modeling approach and its use in understanding weakly connected neural networks see (Hoppensteadt and Izhikevich, 1997, 111).

dating the role of abstraction techniques and the canonical model in explaining a universal neural behavior.

### 3.2.1 Kaplan and Craver’s Mechanist Account

The canonical model approach contrasts with Kaplan and Craver’s claims, because it is used to explain the shared behavior of neural systems without revealing their underlying causal mechanical structure. As the neural systems that share this behavior consist of differing causal mechanisms—different types of ion channels, with different distributions along the membrane, and permeabilities to specific ions, etc.—a mechanistic model that represented the causal structure of any single neural system would no longer represent the entire class of systems with this behavior. Explaining this shared behavior requires abstracting from the mechanistic details of these systems, a feature of the canonical model explanation that conflicts with Kaplan and Craver’s claim that explanations improve with further inclusion of mechanistic detail. A mechanistic explanation can be provided to explain why any *single* system displays class I excitability, but this answers a different question than that answered by the canonical model, which takes the shared behavior of all systems in the class as the desired explanandum. This explanandum does not fit well with the mechanist framework. As Kaplan and Craver state:

it is merely suggestive to note that a similar pattern is observed in a variety of other systems. This information might be useful in our search for general patterns in the organization of mechanisms, but it does nothing to explain the phenomenon we want to explain in the first place. If anything, it merely points out that many other similar phenomena require explanations as well, and perhaps these explanations will be similar.

In contrast with this claim, the canonical model approach is a case where a similar pattern or universal behavior is exactly what neuroscientists want to explain in the first place. Furthermore, dynamical systems neuroscientists do not explain this shared behavior by referencing the causal mechanisms that underlie the neural systems, because their underlying mechanisms differ too greatly for such an approach. Instead, they explain this shared behavior by using a dynamical model that abstracts from mechanistic details and maps onto neural systems in a more complex fashion than the 3M criteria specifies.

These points can be made more clear by considering Kaplan and Craver’s 3M constraint for explanatory models in neuroscience, which the canonical model does not meet. Recall that the first part of this constraint requires that the variables of a model map onto the mechanism of interest, i.e. the entities, activities, and organizational features of the target system producing the phenomena of interest. The canonical model contains a single variable  $\theta$  and the bifurcation parameter  $r$ , representing the behavior of the neuron and a fixed input to the neuron, respectively. The bifurcation parameter does not represent a component (or internal aspect) of the neural system, but rather an input stimulation to the system. This leaves the variable  $\theta$  as a candidate for the first part of the 3M constraint. This single variable ( $\theta$ ) cannot fulfill this constraint because it does not map onto any identifiable entity, activity, or organizational feature of the mechanisms that underlie these neural systems. Rather it represents the overall behavior of the neural system by indicating its location on the unit circle. The second 3M requirement—that variables in the model map onto causal relations in the target system—is also problematic. As the only candidates for a dependency relation in this model are  $\theta$  and  $r$ , it may be claimed that they meet the second part of the 3M constraint: a dependency relation between a fixed input to the neuron and its behavior. However, the fact that these variables do not meet the first part of the 3M requirement makes this dependency relationship difficult to interpret with the mechanist framework. Furthermore, Craver considers this type of input/output relation to be a “phenomenal model” that “black boxes” the underlying causal mechanism. He claims that such phenomenal models are not explanatory because they fail to represent the mechanism between the input and output relations. As he has stated, phenomenal models “are complete black boxes; they reveal nothing about the underlying mechanisms and so merely ‘save the phenomenon’ to be explained” (Craver, 2006, 357). Thus the only possible dependency relation in the canonical model fails to meet 3M because it merely captures an input/output relation and fails to map onto an underlying causal structure.

Indicating that the canonical model does not meet the 3M constraint is not to say that the model does not represent or map onto neural systems in a manner relevant to its explanatory power. Surely models must bear some relationship to how things are in the real world in order to be explanatory. I am not arguing against there being an explanatorily relevant

sense in which the canonical model maps onto physical systems. Instead I am arguing that the mechanists' 3M requirement does not accurately characterize this mapping relationship for all explanatory models in neuroscience. There does not seem to be any straightforward modification of 3M that would allow the mechanist to accommodate the complex relationship between the canonical model and systems with this type of shared behavior.

On this basis it is fair to conclude that the canonical model for class I excitably cannot be accommodated by Kaplan and Craver's mechanist account. This model fails to meet their 3M criteria, their claims regarding the inclusion of details in explanatory models, and their assertion that explanatory models reveal the structure of mechanisms. The specific example that I have provided indicates that even if the mechanist framework accounts for many explanations in neuroscience, it cannot not account for all of them.

### **3.2.2 Batterman's Minimal Model Explanations**

An account of explanation that accommodates this canonical model example is Batterman's account of minimal model explanations. Explanations in science are often considered answers to why-questions and Batterman has distinguished between two different types of these questions: type (i) and type (ii) why-questions ([Batterman, 2001](#), 23). A type (i) why-question asks why a phenomenon manifests in a particular circumstance, while a type (ii) why-question asks why a phenomenon manifests generally or in a number of different circumstances. For example, a type (i) why-question might ask why a particular firing behavior is exhibited by a rat hippocampal neuron. An answer to this question is likely to provide an account of how components of the rat hippocampal neuron bring about the spiking behavior of interest. A type (ii) why-question, on the other hand, might ask why a particular firing pattern is found generally among a group of microstructurally distinct neurons, e.g. rat hippocampal neurons, crustacean motor neurons, and human cortical neurons. An answer to this question is unlikely to reference the lower-level components of the systems, because the components vary from system to system. An explanation for why all of these neurons exhibit the same firing behavior should explain why one can abstract away from the details of each system to achieve the same higher-level behavior. Whenever the lower-level components of

a single system are invoked, explanation of the shared behavior of all these systems is lost.

While mechanistic explanations provide answers to type (i) why-questions, Batterman’s minimal model explanations aim to answer type (ii) why-questions. The first step in these explanations is the identification of a pattern or behavior that is shared among physically distinct systems. This shared behavior is often referred to as universal behavior and the group of systems that exhibit it as the universality class. The universality class can be delimited and made precise by using mathematical abstraction techniques to show how different physical systems display the same universal behavior. Batterman describes this strategy as involving an abstract space of possible systems, where each point in the space represents a particular physical system of interest.<sup>15</sup> The goal is to apply simplifying techniques to this space that allow for the elimination of details or degrees of freedom, while preserving the form of behavior of each system in the space. Repeated application of these techniques (which involve the renormalization group theory in Batterman’s example) rescales the systems and changes their representation in a way that can be tracked as the movement of the system through this space. Studying the topological features of this abstract space reveal fixed points, or points in the space where many represented systems flow to and remain. Importantly, the systems in this space that flow to the same fixed point are in the same universality class and their shared behavior is determined by the fixed point that they all flow to. This procedure of creating, simplifying, and studying systems in this abstract space provides a precise way of delimiting the universality class (Batterman and Rice, 2014). This strategy of delimiting a universality class explains why physically distinct systems all share the same behavior because it reveals that when details irrelevant to the behavior of each system are removed from the models that represent them, all systems share a common representation. As Batterman states:

explanation of universal behavior involves the elucidation of principled reasons for bracketing (or setting aside as “explanatory noise”) many of the microscopic details that genuinely distinguish one system from another. In other words, it is a method for extracting just those features of systems, viewed macroscopically, that are stable under perturbation of their microscopic details (Batterman, 2001, 43).

---

<sup>15</sup>For more on Batterman’s discussion of these points, see (Batterman, 2001, 2010; Batterman and Rice, 2014).

Explaining this universal behavior answers a type (ii) why-question in explaining why physically distinct systems exhibit the same behavior.

Delimiting the universality class can be used to identify what Batterman calls a “minimal model,” which is known to be in the universality class and thus, shares features of all models in the class. A minimal model often provides a compact characterization of universal behavior and, as Nigel Goldenfeld states, is a model that “most economically caricatures the essential physics” (Goldenfeld, Martin, and Oono, 1989; Batterman, 2002). Thus, minimal models characterize the behavior of a universality class without representing the lower-level physical details of systems in the class. Such simple models are often used to study and explain universal behaviors, which Batterman refers to as minimal model explanations. What justifies the use of a minimal model in studying and explaining universal features? This justification is provided by the mathematical techniques that delimit the universality class and the identification of the minimal model as a member of this class.

There are striking similarities between Batterman’s account of minimal model explanations and the explanations provided by the canonical model approach. Like minimal model explanations, the canonical model approach is used to explain the universal behavior of class I neurons. It provides an answer to a type (ii) why-question by explaining why a particular neural behavior is found among physically distinct neural systems. Models of these systems are represented in the abstract space of phase diagrams where mathematical techniques are used to identify the stable features of these models. As Hoppensteadt and Izhikevich write:

“instead of saying that the [canonical] model loses information about the original phenomena, we say that our model is insensitive to the dynamics within an equivalence class...and that it captures properties [of models in the family] that are transversal to the partitioning” (Hoppensteadt and Izhikevich, 1997, 116).

The canonical model for class I excitability is a minimal model in the sense that it provides a compact characterization of the behavior of a universality class, which has been precisely demarcated and includes the canonical model as a member. As Hoppensteadt and Izhikevich state:

Canonical [m]odels arise when one studies critical regimes, such as bifurcations in brain dynamics. It is often the case that general systems at a critical regime can be transformed

by a suitable change of variables to a canonical model that is usually simpler, but that captures the essence of the regime (Hoppensteadt and Izhikevich, 1997, 4).

Moreover:

Using comprehensive models [which attempt to take into account all known neurophysiological facts and data] could become a trap, since the more neurophysiological facts are taken into consideration during the construction of the model, the more sophisticated and complex the model becomes. As a result, such a model can quickly come to a point beyond reasonable analysis even with the help of a computer. Moreover, the model is still far from being complete (Hoppensteadt and Izhikevich, 1997, 3, 5).

Mathematical neuroscientists abstract away from the physical differences among systems that exhibit class I excitability, in order to explain this shared behavior. This procedure involves extracting such behavior with mathematical reduction techniques and representing it with dynamical models. The dynamical models that concisely capture these shared behaviors are often referred to as canonical models. Neuroscientists consider the canonical model for class I excitability a “one-dimensional caricature of a ‘real’ neuron” and they use it to study and explain this universal neural behavior (Gutkin and Ermentrout, 1998).<sup>16</sup>

An all too common objection to the explanatory status of dynamical models has been the claim that—in the absence of representing components of biological mechanisms—they are merely phenomenological models that are only capable of describing or predicting scientific phenomena. Kaplan and Craver insist that “there is no currently available and philosophically tenable sense of ‘explanation’ according to which such models explain,” arguing that their mechanist theory alone best represents the standards of neuroscience. (Kaplan and Craver, 2011, 602). This chapter is intended to refute such claims in light of Batterman’s

---

<sup>16</sup>In a recent paper, Silberstein and Chemero (2013) argue that some explanations in neuroscience fail to meet the mechanistic requirements of decomposition and localization. They state that in these non-mechanistic explanations, “the essential explanatory work is not being done by localization and decomposition.. [but]...the explanatory work in these models is being done by their graphical/network properties and the dynamics thereupon” (Silberstein and Chemero, 2013, 960). My argument for the explanatory status of dynamical models does not require that these models fail decomposition and localization. Rather, I argue that the mechanist position fails to accommodate the explanations I discuss, because it fails to capture the role of abstraction in explaining universal behavior. I take this failure of the mechanist position to be what relates Batterman’s work to the example I discuss. Silberstein and Chemero’s paper does not discuss the relationship between abstraction and the failure of decomposition and localization, but this seems worth exploring in future work. Additionally, I do not argue that dynamical models explain in virtue of their “graphical/network properties and...dynamics,” but rather in virtue of abstracting away from details of neural systems that are irrelevant to a behavior of interest.



account of minimal model explanations and the similarity of this explanatory structure to explanations neuroscientists provide with the canonical model approach. This approach demonstrates how the techniques of dynamical systems neuroscience are used to explain *why* such universal behaviors are exhibited by physically distinct systems, as opposed to just providing descriptions or predictions of these behaviors or revealing their underlying causal mechanisms. Such explanations are provided by simplifying neural models of these systems in a way that reveals their shared qualitative features. That such features are represented by the canonical model is explained by using techniques to demarcate the universality class, of which the canonical model is a member.

I have indicated why Kaplan and Craver’s mechanist position cannot account for the explanations provided by the canonical model approach and how they can be characterized by Batterman’s account of minimal model explanations. This analysis indicates that there are explanations in neuroscience that do not meet Kaplan and Craver’s mechanistic account of explanation and, thus, that it should not be considered the sole form of explanation in neuroscience.

### 3.3 CONCLUSION

Models that are viewed as explanatory all seem to bear some relation to how things are in the real world. In the context of dynamical systems neuroscience, much more can be said about the constraints that explanation places on this relationship and how such constraints are justified. The canonical model approach indicates that this relationship can be much more complex than a direct mapping from variables and their interdependencies, to components and their causal relations. Furthermore, it demonstrates that clarifying these issues involves attending to the specific phenomena scientists aim to explain and the techniques common to their field, as these are likely to influence their approach to understand such phenomena. In a paper involving a dynamical systems analysis of neural excitability, Rinzel and Ermentrout conclude with the following remark:

[W]e emphasize the value of using idealized, but biophysically reasonable, models in order

to capture the essence of system behavior. If models are more detailed than necessary, identification of critical elements is often obscured by too many possibilities. On the other hand, if justified by adequate biophysical data, more detailed models are valuable for quantitative comparison with experiments. The modeler should be mindful and appreciative of these two different approaches: which one is chosen depends on the types of questions being asked and how much is known about the underlying physiology ([Rinzel and Ermentrout, 1989](#)).

The canonical model approach clarifies a type of question that neuroscientists address, but which has received little attention in this philosophical literature. Understanding the approach dynamical systems neuroscientists take in explaining this neural behavior requires attending to their explanandum of interest and the unique modeling tools common in their field. However, such explananda and their respective explanations do not fit well with the dominant mechanist account of explanation. Analysis of the practice and techniques of dynamical systems neuroscience reveals that there are alternative patterns of explanation in this domain.

## 4.0 RECONSIDERING THE MULTIPLE REALIZABILITY ARGUMENT AGAINST EXPLANATORY REDUCTION

### 4.1 INTRODUCTION

While multiple realizability was originally viewed as a problem for theory reduction, it has more recently been discussed as a problem for explanatory reduction. If explanatory reduction requires that explanations appeal to lower level details, then cases where higher level phenomena are multiple realized by different lower level details might seem problematic. Can we appeal to lower level details in explaining some multiply realized higher level phenomena? If so, which details can (or should) we appeal to? Putnam and Fodor suggest that in these cases there is an important sense in which lower level details are irrelevant to and limited in explaining higher level phenomena (Putnam, 1975; Fodor, 1974). Higher level details and generalizations provide better explanations of these phenomena because their broad scope captures unified higher level phenomena that appear disunified at lower levels. In the context of mental phenomena and lower level neuroscience, they suggest that this results in an autonomy of mental phenomena from the lower-level details of the brain. These points support the view that situations of multiple realizability prevent reductive explanation and require non-reductive explanatory approaches.

Many view this anti-reductionist position as unpersuasive and there is continued resistance toward claims that multiply realizability poses any such problems. Sober (1999) and Waters (1990) argue against the claim that multiple realizability prevents explanatory reduction in the biological sciences. They claim that the multiple realization of some higher level phenomena by differing lower level details does not render such details problematic for, or irrelevant to, explanations of these phenomena. They view appeal to lower level realizers

as having various explanatory advantages and they maintain that scientists routinely cite such details in providing explanations. According to them the choice to include or omit lower level details in these explanations is a “matter of taste” with neither option being “objectively preferable” (Sober, 1999; Waters, 1990). Including more detail enhances explanatory depth and including fewer can provide explanatory breadth, while both are valued by scientists. They claim that it is incorrect to view unified higher level generalizations as explanatory privileged, because they have no advantage over reductive explanation in cases of multiple realizability and there are no principled grounds on which their unificatory character provides explanatory strength. Considerable work in philosophy of biology and neuroscience supports Sober’s general position and his refutation of the multiple realizability argument against reduction (Waters, 1990; Bechtel and Mundale, 1999; Bickle, 2006; Butterfield, 2011)

In the context of biology and neuroscience these positions raise a number of puzzles regarding the interaction between multiple realizability and explanatory reduction. If multiple realizability is common in these sciences then Putnam’s position conflicts with the fact that many explanations in these domains do not appear autonomous in the strong sense he argues for. Higher level traits and behavioral phenotypes are often explained by appealing to lower level molecular, biochemical, and neurological details, as opposed to only citing higher level factors. In fact, the success of these explanations in neuroscience, despite the seemingly ubiquitous nature of multiple realizability, has been viewed as an indication that multiple realizability must not pose any real challenge for reductive explanation in this domain (Bechtel and Mundale, 1999). Despite these points, explanations in these sciences also do not appear as boundlessly (infinitely) reductive in the sense supported by Sober and Waters. Explanations in these sciences appear to involve some principled constraints on exactly which lower level details matter, as opposed to being determined solely on the basis of personal tastes or preferences. If some form of reductive explanation is common in these sciences, despite rampant multiple realizability, is it correct to view multiple realizability as problematic for reductive explanation? If so, how do such problems actually manifest in scientific practice? Is there a more nuanced way to characterize the interaction between multiple realizability and reductive explanation that avoids issues associated with the two sides of this debate?

In this chapter, I argue that a particular type of multiple realizability can prevent reduc-

tive explanation in biology and neuroscience, but that it does not imply the strong form of autonomy that some supporters of this thesis advocate. A key feature of my analysis involves the view that some of these issues can be clarified by examining multiple realizability in the context of causal explanation. I examine a particular type of multiple realizability where for some higher level causal relationship the cause is multiply realized by distinct lower level factors that are each (individually) capable of causing instances of the higher level effect. This case is similar to situations of causal heterogeneity, which are common in biology and neuroscience, where there are heterogeneous lower level causes for some higher level type phenomenon. I argue that these situations involve serious problems for appealing to lower level causal factors in explaining higher level type phenomena. I examine a particular explanatory strategy scientists use to circumvent problems found in these situations, where they identify and appeal to a common causal process. I examine the implications of these cases and the explanatory techniques scientists employ for reductive explanation in biology and neuroscience. I pay particular attention to explanation of disease phenotypes in these fields.

## **4.2 THE MULTIPLE REALIZABLY ARGUMENT AGAINST REDUCTION: BOTH SIDES OF THE DEBATE**

In this section I first briefly discuss the main points of Putnam and Fodor's multiple realizability thesis against reduction. I focus on claims they have made in the context of mental, psychological and neural phenomena. I then describe how Sober and Waters respond to this thesis with their focus on the biological sciences. I introduce Sober's example of the causal relationship between cigarette smoking and lung cancer, as this figures significantly in his argument and will be the focus of much of my analysis in this chapter.

### **4.2.1 The multiple realizability thesis**

Some of the earliest and most well-known multiple realizability arguments against reduction are found in the work of Putnam (1975) and Fodor (1974, 1975, 1997).<sup>1</sup> They focus on

---

<sup>1</sup> (Fodor, 1975, 1974, 1997)

higher level phenomena in the special sciences and their relationship to scientific phenomena at the lower levels of molecular biology, chemistry, and physics. As scientific phenomena at all levels factor into laws or causal generalizations they consider how generalizations in the higher level sciences relate to those in the lower level sciences. Fodor considers how a higher level generalization like “if P then Q” or “P causes Q” relates or reduces to lower level scientific generalizations (Fodor, 1975). To use an example discussed by Sober we might ask how the higher level generalization “smoking causes lung cancer” relates to lower level generalizations and phenomena. Where P represents “smoking” (cigarettes) and Q represents “lung cancer” there are multiple lower level realizations of each of these variables, as represented in Figure 12. The different micro-constituent realizations of P include different carcinogenic chemicals in cigarette smoke ( $A_1, A_2 \dots A_n$ ) and the different micro-constituents realizations of Q include different types of cancerous tissues ( $B_1, B_2 \dots B_n$ ). A higher level causal generalization relates P and Q just as lower level generalizations relate each of the cause realizers to their specific effect realizer. The higher level phenomena are instantiated by different micro-level details in a way that prevents them from corresponding in a one-to-one fashion with any specific lower level phenomena. Putnam and Fodor’s multiple realizability argument against reduction focuses on this failure of correspondence.

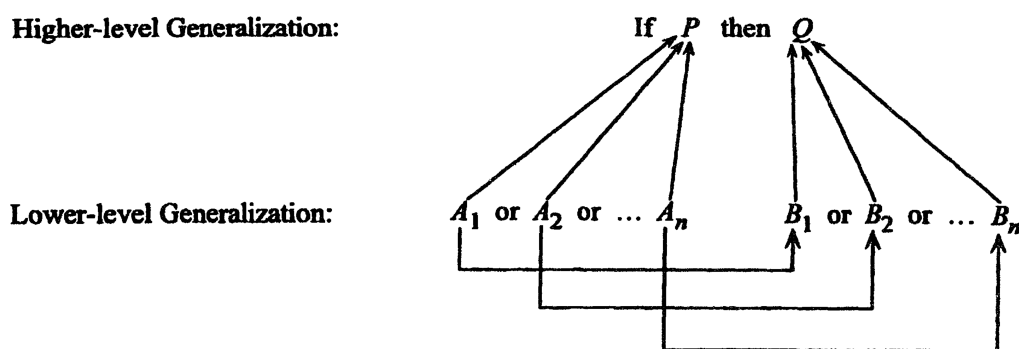


Figure 12: Multiple realizability (Fodor, 1975)

In clarifying the problem multiple realizability poses for reduction Putnam and Fodor distinguish between two reductive claims. The first weak reductive claim maintains that *token* or single instances of higher level phenomena reduce (or correspond) to specific lower level

phenomena. Putnam refers to this claim as the “first thesis” of materialism (Putnam, 1975, 128), Fodor calls it “token physicalism” (Fodor, 1974, 100) and both find it unobjectionable. The second stronger reductive claim maintains that higher level *type* or multiple instances of higher level phenomena reduce (or correspond) to specific lower-level phenomena. Putnam refers to this claim as the “second thesis” of materialism (Putnam, 1975, 128), Fodor calls it “type physicalism” (Fodor, 1974, 100) and both claim that it is untenable on the basis of multiple realizability.

According to Putnam and Fodor multiple realizability is a problem for the second stronger sense of reduction and not for the first weaker sense. The multiple realization of the higher level type-phenomena by lower level phenomena prevents a one-to-one correspondence or reduction of the higher level phenomenon to any lower level phenomenon. The higher level type phenomenon unifies the disjunctive set of lower level realizers in a way that is lost with focus at the lower level and in a way that plays an important role in scientific explanation. As Fodor states, “Even if (token) psychological events are (token) neurological events, it does not follow that the natural kind predicates of psychology are co-extensive with the natural kind predicates of any other discipline (including physics). That is, the assumption that every psychological event is a physical event does not guaranty that physics (or, *a fortiori*, any discipline more general than psychology) can provide an appropriate vocabulary for psychological theories” (Fodor, 1974, 105). This lack of a one-to-one correspondence between higher and lower level phenomena means that the lower level sciences cannot posit the phenomena that the higher level sciences require (Fodor, 1974, 113). According to Putnam, if we try to explain these higher level phenomena by appealing to lower level details we get either no explanation or a “terrible explanation” because the lower level details are simply not relevant to this phenomena (Putnam, 1975). As he states “the same explanation will go in any world (whatever the microstructure) in which those *higher level structural features* are present. In that sense *this explanation is autonomous*” (Putnam, 1975, 131). The superiority of higher level generalizations is often attributed to their broad scope in the sense that they pertain to many systems that differ in their lower level detail (and unify such phenomena that appear disunified at lower levels). This is viewed as a generality that a lower level explanation cannot achieve. For these reasons, cases of multiple realizability where there is

a focus on type-level explananda, are cases where higher level non-reductive explanations are superior to lower level reductive explanations.

#### 4.2.2 Opposition to the multiple realizability argument against reduction

Sober (1999) claims that Putnam and Fodor mistakenly view multiple realizability as problematic for explanatory reduction. He focuses on causal explanation in the biological sciences to refute their claims. Sober refers to the multiple realizability illustration in Figure 12 and distinguishes vertical “synchronic” relations from horizontal “diachronic” relations. Synchronic relations obtain between simultaneously instantiated properties, e.g. between  $P$  and its realizers  $A_1, A_2, \dots, A_n$ , while diachronic relations maintain over time and can be causal, e.g. the relation between  $P$  and  $Q$ , and between  $A_1$  and  $B_1$ . As Sober states:

Since the multiple realizability relation obtains between simultaneously instantiated properties, the [synchronic] relation is not causal...However, the diachronic laws I want to consider *are* causal—they say that a system’s having one property at one time causes it to exhibit another property sometime later. The reason I will focus on causal diachronic laws is not that I think that all diachronic laws are causal, but that these provide the clearest cases of scientific explanation (Sober, 1999, 546).

Sober represents these relations more explicitly with Figure 13, where he considers higher and lower level properties at time  $t_1$ . This figure represents a higher level causal (or diachronic) relationship where  $X$ , at time  $t_1$ , is a cause of  $Y$ , at time  $t_2$ . It also represents a synchronic relationship, between  $X$  and  $X$ ’s realizer,  $Z$ , both at time  $t_1$ . With these clarifications, Sober characterizes the situation, as follows:

“Reductionism says that if  $(x)$  explains  $(y)$ , then  $(z)$  explains  $(y)$ ; it also asserts that  $(z)$  determines  $(x)$ . The multiple realizability argument against reductionism does not deny that higher-level properties are determined by lower-level properties. Rather it aims to [show that]... $(z)$  does not explain  $(y)$ , or so this argument contends” (Sober, 1999, 544).

Thus, according to Sober, the issue at stake is the following: if  $(x)$  explains  $(y)$ , is it also the case the  $(z)$  explains  $(y)$ ? Sober states this more explicitly with regard to Figure 12, in proposing the following question:



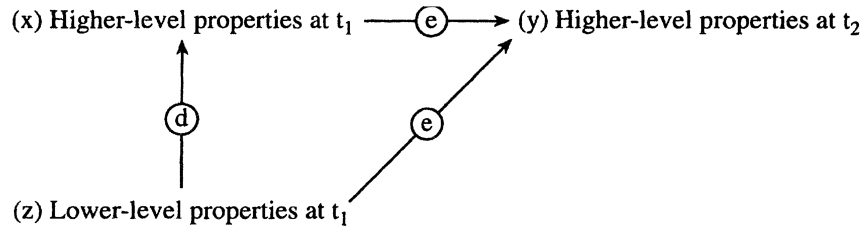


Figure 13: Synchronic (d) and diachronic (e) relations between lower level and higher level phenomena (Sober (1999), p. 544)

- (1) “If an individual’s having property P explains its having property Q, is it also true that its having property  $A_i$  explains its having property Q?” (Sober, 1999, 546).

Sober views the correct answer to be “yes” in line with the reductionist position, while the multiple realizability position denies this claim, incorrectly. Along the lines of this reasoning, if it is correct that P causes Q, and that P is realized by  $A_i$ , then it is also the case that  $A_i$  also causes Q. This follows from the diachronic relationship between P and Q, and the synchronic relationship between P and  $A_i$ .

Sober expands on this position in a number of ways. First, he indicates that, given this reasoning, we are justified in appealing to such lower level details in explaining Q and that this shows how Putnam is wrong to view them as irrelevant to or problematic for such explanations. Waters’ refers to this multiple reliability position as the “Gory Details Argument,” because it mistakenly views the complex lower level details as some kind of problem for reductive explanation (Waters, 1990, 131). Instead, including more detail allows for explanatory “depth” and including less allows for explanatory “breath,” and scientists seek both for purposes of explanation. Both Sober and Waters indicate that appealing to lower level details can have a number of important explanatory advantages. According to Waters, because the “more uniform” higher level perspective provides “shallow” explanations, compared to the “deeper” lower level explanations, there is no sense in which the former are “objectively preferable” to the later. Even if the uniformity of the higher level explanation did provide an explanatory advantage, there is no reason to think that it could not also be

captured at the lower level (Waters, 1990, 131). They claim that Putnam and Fodor misleadingly act as if there is some “objective rule” whereby including fewer or more of these details influences explanatory power, while it is more correctly viewed as “simply a matter of taste—do we want more details or fewer?” (Sober, 1999, 551) As Sober states: “Perhaps the micro-details do not interest *Putnam*, but they may interest *others*, and for perfectly legitimate reasons. Explanations come with different levels of detail. When someone tells you more than you want to hear, this does not mean that what is said fails to be an explanation. There is a difference between explaining too much and not explaining at all” (Sober, 1999, 547; emphasis original).

Another feature of Sober’s argument is that the multiple realizability argument against reduction is inconsistent with scientific practice. Scientists routinely appeal to such details in providing reductive explanations for multiply realized phenomena. If multiple realization was a real problem for reductive explanation, it is unclear how scientists could provide such explanations. Waters’ supports this line of reasoning too, claiming that the “explanatory edge” of the higher level explanation only has appeal when “our attention is called away from the actual biology” (Waters, 1990, 132). Aside from multiple realizability being a non-issue for reductive biological explanation, scientists derive significant explanatory value from appealing to lower level details. Sober refers to this in the context of the aforementioned smoking causes lung cancer examples:

“I very much doubt that the concept of explanatory relevance means what Putnam requires it to mean in this argument. When scientists discover why smoking causes cancer, they are finding out which ingredients in cigarette smoke are carcinogenic. If smoking causes cancer, this is presumably because the micro-configuration of cigarette smoke is doing the work. If there turn out to be several carcinogenic ingredients and different cigarettes contains different ones, this does not make the molecular inquiry explanatorily irrelevant to the question of why people get cancer. The fact that P is multiply realizable does not mean that P’s realizations fail to explain the singular occurrences that P explains. A smoker may not want to hear the gory details, but that does not mean that the details are not explanatory” (Sober, 1999, 548-9).

Additionally, he states:

“How are we to explain why this person has that disease? One possible reply is to say that the person smoked cigarettes. A second possibility is to say that the cancer occurred because the person inhaled ingredient  $A_1$ . Putnam’s multiple realizability argument entails that the

second suggestion is either no explanation at all, or is a “terrible” explanation. I suggest, however, that it should be clear to the unjaundiced eye that the second explanation may have its virtues...The additional details provided by the micro-explanation are not stupid and irrelevant. They make a difference—to the probability of the *explanandum*, and to much else. Perhaps it is a good thing for cancer research that the multiple realizability argument has not won the hearts of oncologists” (Sober, 1999, 548-9).

Many view Sober’s response to the multiple realizability argument against reduction as extremely successful. Butterfield (2011) strongly endorses Sober’s argument and claims that “Sober has definitely refuted this [multiple realizability] argument, in its various versions.” Considerable work in philosophy of biology and neuroscience supports Sober’s general position (Waters, 1990; Bechtel and Mundale, 1999; Bickle, 2006).<sup>2</sup> In the following section I examine the smoking example in more detail to argue that multiple realizability does pose a problem for reductive explanation, which has been overlooked in this debate.

### 4.3 ANALYZING THE SMOKING EXAMPLE: MULTIPLE REALIZABILITY AND CAUSAL HETEROGENEITY

In this section I start with a sketch of the multiple realizability problem for reduction that I will argue for. I further examine the type of multiple realization found in the smoking example and discuss its relationship to situations of causal heterogeneity. I then turn to a more detailed argument for the problems this type of multiple realizability poses for attempts to explain phenomena in biology and neuroscience by citing lower level causal details.

My analysis relies on basic features of Woodward’s interventionist account of causation (Woodward, 2003). The interventionist account is relevant to this debate, in part, because Sober focuses on causal explanation and counterfactual approaches to causation. A significant feature of the interventionist approach is that it views causal relationships as relationships that allow for control and manipulation. On this view causal relationships are characterized by the following interventionist condition: X is a cause of Y, if and only if,

---

<sup>2</sup>Alternatively, Batterman (forthcoming) disagrees with Sober’s claims and identifies various problems with his response (Batterman, ). Batterman’s analysis focuses on non-causal explanation, while my analysis and claims will be restricted to causal explanation in biology and neuroscience.

an ideal intervention on the value of  $X$  (and no other value) were to occur in background conditions  $B_i$ , this would change the value of  $Y$  (Woodward, 2010).  $X$  can be thought of as a handle or switch that allows for control over  $Y$  in the sense that changes in the value of  $X$  control the value of  $Y$ . Such interventions need only be hypothetical in the sense that ideally intervening on and changing  $X$  *would* change  $Y$  without requiring that the intervention is achievable with current technology or in an anthropocentric sense. On this account explaining some phenomenon of interest requires appealing to its causes or its difference makers. Thus, an explanation of  $Y$  in this scenario involves appealing to  $X$ , as it is a factor upon which  $Y$  counterfactually depends. There are many important and deep features of the interventionist account that I will not discuss here.<sup>3</sup> I rely on a basic understanding of this framework and introduce further aspects of it during my analysis when necessary.

### 4.3.1 Initial sketch of the problem

Consider a brief sketch of the problem I will argue for. Recall Sober’s smoking example where smoking ( $P$ ) causes lung cancer ( $Q$ ). The variable representing smoking ( $P$ ) is multiply realized by distinct carcinogenic substances present in cigarette smoke, including  $A_1$ ,  $A_2$ ,  $A_3$ . Suppose we want to explain (a) the occurrence of lung cancer among three individuals, where the first individual smoked cigarettes with substance  $A_1$ , the second with substance  $A_2$ , and the third with substance  $A_3$ . What is the cause of lung cancer in this population? We are unlikely to claim that  $A_1$  caused these cases of lung cancer, because while it did cause the first case it did not cause the other two. Similarly, we are unlikely to appeal to either  $A_2$  or  $A_3$  because neither makes a difference to all cases of lung cancer. If we want to explain (b) lung cancer in the first individual (or in any population who acquired this cancer after only smoking cigarettes with  $A_1$ ) then we will find appealing to the lower level cause  $A_1$  unproblematic, as it is responsible for these cases. However, (a) and (b) represent different explanatory targets where an *explanatory target* is fixed by both the phenomenon of interest and population one wants to explain the phenomenon in. Once the target is fixed providing an explanation involves citing factors that are causally relevant to it. As I will discuss

---

<sup>3</sup>For more details see (Woodward, 2003).

in more detail below, the lower level factor  $A_1$  is not causally relevant to (a) in the same way that it is causally relevant to (b). However, the higher level cause variable P (which represents “smoking cigarettes”) *is* similarly causally relevant to both (a) and (b) as it makes a difference to all cases of lung cancer in the population.

These claims require much more analysis and clarification, but they immediately raise an issue that it will be helpful to note before moving on. Recall Sober’s affirmative response to question (1) and his claim that if an instance of lung cancer (Q) is explained by smoking (P) it is also explained by carcinogenic chemical ( $A_i$ ). He claims that both causal factors explain of “singular occurrences” of lung cancer (Q), as represented in Figure 13. A significant problem with this analysis is that Sober examines explanations of token-level phenomena, while Putnam and Fodor’s argue that multiple realizability is problematic for explanations of type-level phenomena.<sup>4</sup> Focus on type-level phenomena is, in a sense, inherent to the concept of a multiply realizability in that it involves some property that is realized by multiple instances of some kind, as opposed to a single or token instance. Secondly, that the intended target of this thesis is type and not token level phenomena is indicated by Putnam and Fodor’s explicit issues with the “second thesis” of materialism and “type physicalism” on the basis of multiple realizability, and their non-issue with “first thesis” of materialism and “token physicalism,” respectively. A final indication of the mis-match between Sober’s analysis and the multiple realizability thesis, is that those who support this thesis commonly invoke the importance of generalizations or explanations that have broad scope in pertaining to multiple instances of some phenomena or multiple systems that exhibit some phenomena. The feature of pertaining to multiple instances or systems clearly cannot be appreciated by focusing only on single instance explanations. Thus, the issue at stake in this debate is not question (1), but instead: (2) Can one appeal to the lower level carcinogenic realizers of cigarette smoke to explain the *type-level* occurrence of lung cancer? In this section I will argue that there are compelling reasons to view appeal to such lower level factors as problematic and limited in explaining such higher level type phenomena.

---

<sup>4</sup>Batterman (forthcoming) has identified this and other issues with Sober’s analysis.

### 4.3.2 Important features of the smoking example

Let us consider the smoking example in more detail. In the case where there are three lower level realizers of  $P$  we have four main causal relationships that lead to effect  $Q$ : three involving the lower level causes (i)  $A_1 \rightarrow Q$ , (ii)  $A_2 \rightarrow Q$ , (iii)  $A_3 \rightarrow Q$  and one involving the higher level cause (iv)  $P \rightarrow Q$ . I will point out two important features of this example and then discuss how realistic it is. First, this example involves a particular type of multiple realization where some higher level cause variable and its lower level realizers are *all* causally relevant to the same higher level effect. (In other words, with regard to effect  $Q$  all lower level carcinogens meet the interventionist condition for causation.) Of course, the smoking variable  $P$  is multiply realized by many other properties that are not causally relevant to  $Q$ —e.g. cigarette color, aromatics, non-carcinogenic elements, benign micro-contaminants, etc.—and these realizers do not figure into the example. From the perspective of causal explanation we have good reason to view these realizers as less important than the carcinogens for explaining  $Q$ , because they are not causally relevant to  $Q$ . This suggests that cases of multiple realizability that seem uninteresting or unproblematic for reductive explanation may differ from the type of multiple realizability in this example. Recall the objection that multiple realizability is too ubiquitous despite successful scientific explanation for it to be problematic for reductive explanation. This objection may fail to distinguish important types of multiple realizability and the different consequence they can have for reductive explanation.

These points relate to a second important feature of this example. This example involves a particular form of causal complexity where there are non-specific causes or different causal factors ( $A_1, A_2, A_3$ ) that can produce the same type-level effect ( $Q$ ). I refer to this situation as *causal heterogeneity* to highlight the fact that it represents situations where there are different (or heterogeneous) causes across instances of the type-level effect of interest. This situation contrasts with cases where type-level effects have some specific shared causal process, e.g. in cases where disease phenotypes are said to have some shared causal etiology. Figure 14 illustrates the type of multiple realizability and causal heterogeneity found in this example. Causal heterogeneity alone does not specify information about the levels at which the cause or effect variables are located so the smoking example is a particular case

where the heterogeneous causes are at a lower level than the effect they produce. Situations of causal heterogeneity are common in biology and neuroscience. Examples include genetically heterogeneous disease phenotypes like familial hypercholesterolemia (very high LDL cholesterol), Parkinson’s disease, and retinoblastoma pigmentosa (progressive severe vision loss) where completely distinct gene variants cause the same disease phenotype in different individuals. These variants are sometimes called “rare” variants, because they only cause a subset of the population level disease as opposed to being a “common” variants that represent common causes of the disease. Another example is the shared higher level behavioral effect of drug addiction, which can be caused by different chemical drugs, including nicotine, alcohol, and opiates. Thus, the type-level behavioral phenotype of addiction is said to be causally heterogeneous in that it can be caused by different chemical components. There is an important difference between the smoking example and common cases of causal heterogeneity in biology and neuroscience. In many cases of causal heterogeneity there is not always a clear higher level cause variable that represents all lower level realizers. This is seen in the cases of genetically heterogeneous disease phenotypes, because there is no single higher level cause that represents a property that is realized by the gene variants. These situations can be represented by the illustration in figure 14, where the multiple realizability relations (a) are omitted, but the relationships of causal heterogeneity (b) remain. In the smoking example, the standard many-to-one realizer to realized property relation creates a situation where there is a many-to-one relationship between the lower level causes and higher level effect. In this section, I will further examine how these features are related and how they pose problems for explanations of type-level phenomena.<sup>5</sup> I focus on causal heterogeneity in this analysis.

---

<sup>5</sup> A final point to mention is that causal heterogeneity differs from a type of causal complexity for token-level phenomenon. I refer to this type of causal complexity as *multicausality*, where many different causes work in aggregate to produce a single case or token-level effect. These two types of non-specificity are not mutually exclusive. While the smoking example involves single factors that cause each single cases of the disease, there are, of course, many other disease examples, where individual cases of the disease are explained by appealing to multiple causes working together (PKU, gene and diet). These cases involve multicausality. They may also involve causal heterogeneity, if in addition to their being many causes working in aggregate to produce single cases, these groups of causes differ across cases. For example, to say that disease X in patient  $N_1$  is caused by  $C_1$  and  $C_1$  is to say that it is multicausal, in the sense that the single case of the disease involves more than one cause. The same disease may be causally heterogeneous, if disease X is produced in patient  $N_2$  by causes  $C_3$  and  $C_4$ . Thus, causal heterogeneity captures differences in causal factors across cases of the effect, while multicausality captures difference in causal factors within single cases of the effect.

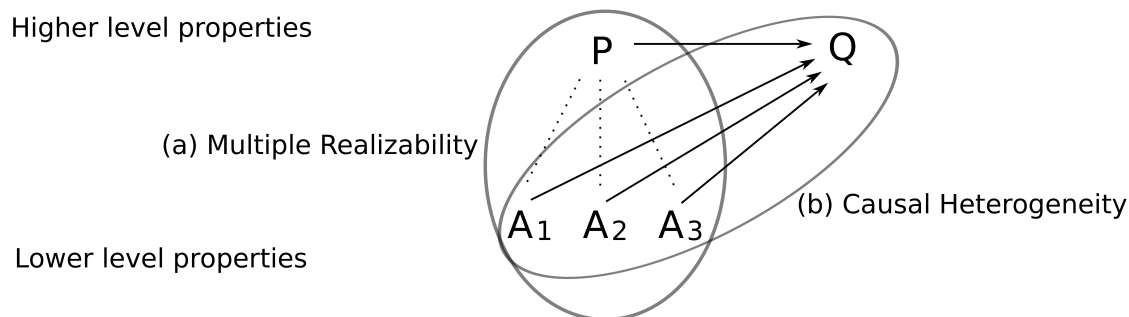


Figure 14: This figure represents the (a) multiple realization and (b) causal heterogeneity found in the smoking example. Relationships of realizations are represented with dashed lines and relationships of causation are represented with solid arrows

How realistic is the smoking example? First, individual cigarettes contain mixtures of many different types and amounts of carcinogenic substances, as opposed to containing just single carcinogenic substances (Hecht, 2003). Thus, instead of having different lower-level factors (or groups of factors) that produce different cases of lung cancer, all cases are produced by a fairly consistent menagerie of carcinogens. The feature of having different lower level causes across instances of the effect is more representative of genetic heterogeneity and other cases of causal heterogeneity where the causes clearly vary across cases of the disease. Second, these points relate to another feature of the smoking, which is the multiple realization of the effect variable “lung cancer” by distinct lower level molecular forms of cancer. Smoking cigarettes can cause different types of primary lung cancer,<sup>6</sup> but the mixture of carcinogens in cigarette smoke makes it unclear whether these types track with particular carcinogens.. What is clear is that the smoking cigarettes and carcinogenic substances they contain causes the single higher level effect of lung cancer, regardless of the specific cellular type of cancer. Finally, smoking is of course not the only cause of lung cancer (other causes include: asbestos, radon gas, and air pollution), but if we reduce our focus to primary lung cancer,<sup>7</sup> which is standard in the biomedical context, it is responsible for the overwhelming

<sup>6</sup>The four main types of lung cancer are classified histologically and include squamous cell carcinoma, adenocarcinoma, large cell carcinoma, and small cell carcinoma (Hecht, 2003, 733).

<sup>7</sup>Primary lung cancer refers to lung cancer that originates in the lung as opposed to metastasizing from



majority of these cases.<sup>8</sup> Despite some of the deviations of the smoking example from its representation in the biomedical context, the fact that the structure of his example well represents other common situations in biology and neuroscience makes further analysis of this example worthwhile. Another point in support of this, is that it takes little modification of the smoking case to view as having the structure represented in Sober’s example. I now examine the smoking example further and turn to situations of causal heterogeneity in the following section.

### 4.3.3 Smoking example: problems for appealing to the lower level carcinogens

With these clarifications in mind, consider a similar situation to the earlier scenario with three individuals. We now want to explain (c) the occurrence (and nonoccurrence) of lung cancer in a larger population, say among the current human population. A portion of this population has lung cancer, while the rest is lung cancer free. Of all cases of this cancer in the population 20% are caused by  $A_1$ , 30% are caused by  $A_2$ , and 50% are caused by  $A_3$ . Recall that the causal relationships between each of these lower level causes and lung cancer (Q) are specified by the three invariant relationships (i), (ii), and (iii). For each of these relationships, suppose that there is some threshold value of the cause variable, such that if it takes a value above this (H= high) it leads to lung cancer (Q= yes) and below this value (L=low) it does not (Q = no). Changes in the value of the cause variable [High, Low] allow for control over the effect variable [Yes, No] and occurrence of the disease. The cause variable can be thought of as a course-grained “switch” that causes cancer when in the “ON” (or High) position and does not cause cancer in the “OFF” (or low) position.

For any individual in this population, this switch-like relationship makes sense of the control that the cause variable has over the effect: if an individual smokes a high level of  $A_1$  she will acquire lung cancer and if she does not (or she just smokes a low level), with no other changes, she will not develop lung cancer. However, explaining lung cancer in a particular individual identifies a *different explanatory target* than asking for an explanation of lung cancer in this population. In this population, different lower level causes matter for

---

some other location.

<sup>8</sup>Cigarette smoking is said to cause 90% of all lung cancer cases ([Hecht, 2003, 733](#)).

different cases of the disease and there is no single specific lower level factor that causes *all* cases of the disease. Furthermore, of all individuals in this population that are cancer free, they are not cancer free just because they refrained from smoking a single carcinogen, but because they refrained from smoking all three of them. The problem that this poses for citing lower level causes is the following: if causal explanation of some phenomena involves citing factors that are causally relevant to this phenomena, the lower level causes do not meet such a requirement in this situation, because there are no specific lower level factors that are causally relevant to *all* cases of the disease that we want to explain. Thus, the problem for reductive explanation in this example is not that the lower level details are “gory,” as Waters suggests, but that the lower level details are *different*. The lower level causal details are different across cases of the phenomena that we want to explain. Appealing to the distinct lower level cause of any one system allows us to explain the effect in that system, but prevents us from explaining the shared effect across systems with different causes.

I will suggest three points that this analysis supports. The first pertains to the role of the explanatory target in providing causal explanations. While it is widely recognized that the causal factors cited in some explanation are constrained by the explanatory target, what has been less appreciated is that this target shifts, not just when the phenomena to be explained changes, but when the population one wants to explain the phenomenon in changes.<sup>9</sup> For any phenomenon or behavior we want to explain in biology or neuroscience the causal factors we appeal to depend on the systems of interest. We might want to explain some phenomenon like aerobic metabolism in different populations like all aerobic organisms, all mammals, a specific species, a sub-group of a species, or a particular tissue or cell type. When explaining aerobic metabolic in a specific species we are likely appeal to and be interested in causal factors that are relevant to metabolism in this species, even if they are not relevant for metabolism in all or most other organisms or systems.<sup>10</sup> Alternatively, if we are interested in explaining

---

<sup>9</sup>I do think that Batterman (2001) has explicitly discussed this point, but perhaps more in the context of non-causal explanation. The point I discuss here is captured in part by Batterman’s distinction between type (i) and (ii) why-questions (Batterman, 2001).

<sup>10</sup>For example, in attempting to understand and explain human diseases, we aren’t necessarily bothered by causal factors that fail to reproduce the disease in non-human animals, unless our explanatory target encompasses this population. Alternatively, subfields of biology that focus on non-humans animals don’t necessary care about whether the causal factors that are relevant for their population are also relevant to the human population.

this phenomena for all aerobic organisms we will appeal to causal processes that “highly conserved” and shared in this population. When the explanatory target shifts from one population to another the factors that were causally homogeneous (or shared) for producing the phenomenon in first population may be heterogeneous causes for the phenomenon in the second. As this causal heterogeneity operates across multiple instances or type-level effects, focusing on explanations of single instances as Sober does, will fail to capture this feature and its influence on causal explanation. Furthermore, any analysis of explanation that aims to be representative of explanatory practice in biology and neuroscience should attend to type-level explanation as it is very common in these sciences.<sup>11</sup>

Second, the interventionist view that causal relationships allow for control helps to clarify the difference between citing lower level and higher level details. Once the explanatory target is fixed on this kind of type level phenomena the problem with appealing to the lower level causes is that they do not allow for control over, and thus they do not account for, the entire explanandum, but only some subset of it. These lower level factors have a many-to-one relationship to the entire explanandum and individual lower level cause only allow for control over and explain a portion of the population-wide phenomenon. This problem can be overcome with the identification of some *shared* cause that accounts for the population level phenomenon, because this reestablishes a the correspondence between a particular causal process and the entire type level effect.<sup>12</sup> This is what the higher level variable “smoking” allows for in this example as it identifies a shared cause of the phenomenon in the population, despite the lower level causal heterogeneity. These claims are further supported by considering the different types of causal control that the higher and lower level factors allow for. Imagine a preventative measure or treatment that targets  $A_1$ , e.g. completely eradicating all  $A_1$  cigarettes from this population or treating all individuals with a drug that immediately nullifies (or cures) all pathogenic effects of substance  $A_1$ . Both of these measures would successfully treat and prevent cases of lung cancer caused by  $A_1$ . However,

---

<sup>11</sup>This is, of course, in addition to my claim that type-level explanation is a more appropriate target of the multiple realizability thesis, than token level explanation.

<sup>12</sup>This suggests an interesting similarity to Putnam and Fodor’s claims of the lack of correspondence between higher level phenomena and their multiple realizations. In this case we have a lack of correspondence between a higher level effect and its multiple (or heterogeneous) causes. The correspondence is reestablished by the identification and appeal to a shared causal process for the type level phenomenon.

without targeting the other causes of lung cancer these measures are limited in controlling the disease at the population level. If a combined total of 80% of lung cancer cases are caused by  $A_2$  and  $A_3$ , then these cases of the disease are not influenced by treatment measures that focus on  $A_1$ . For this population,  $A_1$  is not a causal variable that can be targeted to control all cases of the disease, although it can be targeted to prevent some of them. Substances  $A_2$  and  $A_3$  are also limited in the sense that each cause only certain cases of the population level effect, although they each account for a larger number of cases than  $A_1$ . However, where the lower level carcinogens allow for control over subsets of the population level phenomenon of lung cancer, higher level variable “smoking” allows for control over the phenomenon in the entire population regardless of the lower level carcinogens. This makes sense of the fact that we do view smoking as the cause of lung cancer in humans, instead of attributing the disease to some particular lower level carcinogen. It also makes sense of the fact that the medical community focuses on this higher level cause in explaining, preventing, and treating the disease at the population level. Of course lower level details figure into explanatory and treatment practices when the focus is on individual patients. This makes sense given that it involves a different explanatory target than a population-wide considerations. The different explanatory targets in these sciences constrain which causal factors are cited and the ways in which they are relevant for such phenomena.

Appreciating that there are type-level explananda in addition to token-level explananda still leaves us with a complication. With regard to the type-level phenomenon of lung cancer the lower level carcinogens do not seem *completely* irrelevant as Putnam’s claims appear to suggest. Identifying carcinogen  $A_1$  does allow us to explain and potentially treat some cases of lung cancer in this population. However, citing  $A_1$  does also seem limited in explaining lung cancer in this population in a way that Sober’s analysis fails to appreciate. There seem to be good reasons for viewing  $A_1$  as an incomplete or insufficient explanation of this phenomenon (on the basis of more than ‘matters of taste’). How can we make sense of the more nuanced way in which the lower level carcinogens are (or are not) causally relevant to this phenomenon? The third and final point I will suggest here, is that this can be done with a recharacterization of the notion of “scope” that is consistent with the interventionist framework and my clarification of an explanatory target.

In the extant literature scope is commonly defined without specifying an explanatory target and by assessing the total number of systems that a model pertains to given *all* current or actual systems in the world or some unspecified large number of systems. Such characterizations of scope are found in debates about whether models that have a larger or narrower scope (or are more or less general) are more or less explanatory. This is seen in the back and forth debate between claims, like those made by Putnam who states that “an explanation is superior if it is more general” and claims made by Waters who suggests that models with this generality provide “shallow explanations” compared to the “deep explanations” afforded by less general, reductive approaches (Putnam, 1975, 132) (Waters, 1990, 131). Although each side argues for the explanatory importance of models with large or narrow scope, the debate leads to a standstill because without considering an explanatory target the scope of a model does not consistently track its explanatory value. It should be unsurprising that defining scope in a way that is disconnected from the explanatory target and context will lead to viewing scope as irrelevant to explanatory status. What is more troubling is that such a definition seems unlikely to capture a notion of scope that is germane to biology and neuroscience as the applicability of explanatory models in these sciences is rarely assessed by considering the applicability of models to all systems (or a very large numbers of system) in the world. Instead, judgments of the scope of an explanatory model in these domains are more often made on the basis of the number of systems the model pertains to given some explanatory target that identifies a particular population of interest.<sup>13</sup>

When the notion of scope is restricted to an explanatory target this clarifies: (1) the different ways in which the higher level and lower level causes are relevant to the phenomenon in the smoking example, and (2) the role scope can play in causal explanation and why one model may be viewed as more or less explanatory on the basis of being “more general” or having a larger scope of applicability.<sup>14</sup> On this view, the scope or generality of a causal model

---

<sup>13</sup>For example, see: (Woodward and Hitchcock, 2003; Hitchcock and Woodward, 2003). I think that this notion of scope has interesting connections to the development and popularity of the DN model of explanation. Furthermore, this notion of scope is very similar to claims that a models are “general” or allow for some kind of “unification.” While I find these worth examining in more detail, I do not address them further as they are outside the topic of this chapter.

<sup>14</sup>I do not deny that it may also be useful or helpful to have a notion of scope that is unfixed by a particular explanatory target.

is determined by the number (or percentage) of instances or systems the model accounts for *given some explanatory target*. For the explanatory models in the smoking example (i, ii, iii, and iv) we can view the scope of each model as the portion of systems in the population (c) that the models pertain to, in the sense the invariant relationship and causal variables in the model represent the systems and can be cited in explaining the effect in the system.<sup>15</sup> The invariant relationships specifying the lower level causes—i.e. the explanatory models (i), (ii), and (iii)—each pertain to subgroups of the population level effect, while the invariant relationship specifying the higher level cause (iv) pertains to all cases of the effect in the population. On this view it makes sense to view the lower level causal models (i, ii, iii) as having a narrow scope compared to the higher level causal model (iv). The scope of these models relates to their differences in explaining the type level phenomenon on the basis of the control they allow for. If we want to explain a type level phenomenon it makes sense to view causal models with broad scope as providing better explanations than those of narrow scope, because they identify variables that allow for control over the effect in more cases in the population of interest.

#### 4.3.4 Objections

I will address two objections to this analysis before further examining its consequences for explanatory reduction. In contrast to my claims, it may be argued that the lower level heterogeneous causes *can* be cited in explaining the population level effect if either (1) the higher level phenomena is divided or ‘split’ on the basis of the lower level causes, or (2) the higher level phenomena is kept ‘lumped,’ while one appeals to the disjunctive set of lower level causes. With regard to the first objection it may be claimed that the invariant relationships (i), (ii), and (iii) are not the right ones to consider, because the lower level causes should not be related to the higher level effect, but instead to the lower level instantiations of the effect. These invariant relationships are: (i\*)  $A_1 \rightarrow B_1$ , (ii\*)  $A_2 \rightarrow B_2$ , and (iii\*)  $A_3 \rightarrow B_3$ , as suggested by Figure 13. Along the lines of this objection it may be argued that the higher level variable “lung cancer” does not represent a single phenomena, because it has

---

<sup>15</sup>This is similar to Woodward’s characterization of scope, with the exception of restricting it to an explanatory target of interest (Woodward, 2003, 269).

heterogeneous causes or because it has different lower level molecular realizations. Once the lumped phenomena (Q) it is divided into appropriate categories ( $B_1$ ,  $B_2$ ,  $B_3$ ) we no longer have the problems associated with causal heterogeneity. Now each cause does act as a reliable ON/OFF switch for the effect, even at the population level, and it has control over the effect in the sense that the effect counterfactually depends on the cause in question.

The first thing to note about this objection is that dividing up the higher level effect identifies a different explanatory target than the one I examine in my analysis, which is a type-level effect with heterogeneous lower level causes. One of my aims has been to show how these explanatory targets differ and how the problems I identify arise in the latter type of situation, without necessarily presenting in the former. To focus on explaining only single instances of lung cancer or cases caused by a single common carcinogen is not to engage with or suggest a solution to the problems I discuss. Instead, these focus on explanatory targets where such problems do not occur. This objection may be followed up with the stronger claim that the type level phenomena that I consider are not appropriate explanatory targets. Thus, this claim does not ignore the explanatory target in my analysis, it argues that there are reasons to view it as inappropriate on the basis of grouping disparate phenomena. This objection suggests replacing such a ‘lumped’ phenomenon with various phenomena that track with the relevant distinctions. I do think it is right that scientists often view it as appropriate to “divide up” a phenomena of interest when their evidence suggests that it is unavoidably causally heterogeneous, in the sense of lacking some shared causal process.<sup>16</sup> I discuss this situation more in the next section. In the smoking example, however, dividing up the effect can be avoided on these grounds because the higher level cause “smoking” is a shared cause of the higher level effect. Furthermore, an indication that the explanatory target is appropriate by the standards of scientific practice is that it is a target that we often identify and explain in the biomedical context.

The second objection maintains that the higher level effect can be explained by appealing to the lower level causes, this just requires appealing to a disjunctive set of these causes. A

---

<sup>16</sup>I think that this is often, but not always, the case when they seek a causal explanation of some shared higher level behavior. Of course, they may also seek to provide a non-causal explanation of such phenomena, in which case they are likely use other explanatory strategies (than the identification of shared causal etiology), i.e. minimal model or other explanations ([Batterman, 2002](#)).

problem with this objection is that disjunctive explanations do not appear to well represent scientific explanations in these domains. In these scientific fields there is a strong tendency to isolate and appeal to shared causal processes for disease phenomena and the approach of appealing to sets of disjunctive causes seems very rare (or not present at all). Consider explanations of infectious diseases like anthrax, tuberculosis, and cholera. These diseases are explained by appealing to specific shared contagions, viz. anthrax bacilli, tubercle bacilli, and comma bacilli, respectively. This objection suggests that it is appropriate (or ideal) to appeal to the disjunctive set of these contagions in explaining a particular disease phenotype like “anthrax-tuberculosis-cholera.” However, this disjunctive move does not appear either descriptively or normatively accurate of these scientific explanations. The appeal to shared causal etiology makes normative sense because it is useful in allowing for control over the population-wide occurrence of disease, in a way that a disjunctive explanation is not. While intervening on any one of the disjunctive causes in causally heterogeneous populations can allow for control over a portion of the population, but it cannot not allow for population-wide control.

In this section I have argued that appealing to the lower level carcinogens is seriously limited and problematic for explaining the type level phenomenon of lung cancer in a causally heterogeneous population. I have suggested that appealing to the higher level variable “smoking” does not have these same problems. My analysis suggests that the identification of some shared causal process is important for explaining type level phenomena in biology and neuroscience. I should clarify this suggestion in two ways before proceeding. First, this should not be interpreted as the claim that it is important to identify some single token causal factor for an effect, as we have with the single cause variable “smoking” and as is often intended in discussions of “monocausal” models of disease. The shared causal process I am referring to involves the *same* or *shared* causal factor(s) across cases of a disease, as opposed to a *single* token causal factor for any number of disease cases. Thus, my focus is on a disease phenotype having the same causal process in a population, as opposed to having a single token cause that can be represented by a single variable or monocausal model.<sup>17</sup> The second

---

<sup>17</sup>For example, in the case of the disease phenylketouria (PKU) a gene variant and dietary substance work together in aggregate to produce the disease. As there are two important causal factors in this example, it does not meet a standard interpretation of the monocausal model. However, as these two causal factors are



point is that I am not arguing that all type level phenomenon in biology and neuroscience do have or should have shared causal processes. However, in the context of disease phenotypes in these domains, which I focus on in this chapter, they often do. This is a result of the fact that such phenotypes are defined and classified in a way that is based on their (1) causes and (2) a focus at the level of the human population. One of my goals in this section was to indicate that Sober's smoking example contains a particular type of multiple realizability that relates to cases of causal heterogeneity. I have not yet drawn any explicit conclusions for how this type of multiple realizability and causal heterogeneity influence the potential for reductive explanation in biology and neuroscience. I now turn to this in the next section.

#### **4.4 CAUSAL HETEROGENEITY: PROBLEMS FOR REDUCTIVE EXPLANATION**

In the previous section I indicated that the smoking example involves a type of multiple realizability that has important relations to causal heterogeneity. Both of these features contribute to a problem for appealing to lower level details in explaining the type level phenomenon of lung cancer. Where the type level phenomenon has different lower level causes across systems in some population of interest, such causes are limited in explaining the population wide phenomenon. I argued that the higher level causal variable avoids these problems because it is a shared causal variable for the type-level effect in the population. This may seem to suggest that for this kind of explanatory target appeal to the higher level cause provides a better explanation than appeal to lower level causal factors in a way that supports the explanatory superiority of the non-reductive, autonomous higher level explanation. There is a complication for straightforwardly exporting this conclusion to explanatory practice in biology and neuroscience. Many situations of causal heterogeneity in this domain lack a clear higher level cause variable that represents or encompasses all lower level heterogeneous causes. Thus, the solution employed in the smoking example (of

---

found in all cases of the disease it does meet the shared causal etiology standard. The point is that my focus is on shared causal etiology, not monocausal etiology.

appealing to the shared higher level cause), is not available for many common situations in this domain. I will indicate that scientists employ a similar strategy of searching for and appealing to shared causal factors in these situations. However, these shared causal factors are often located at any level from the lower levels of the heterogeneous causes to the higher level effect. I examine how this influences the potential for reductive explanation in these situations.

What does it mean to provide a reductive explanation of some phenomenon in biology or neuroscience? Three views that may be distinguished include the claims that reductive explanation in this domain requires citing causal details that are (1) at any lower level than the effect, (2) the lowest levels of biology, e.g. the levels of biochemistry or molecular biology, or (3) the lowest levels of physics. Early discussions of theory reduction mainly focused on the third strict sense of reduction. More recent discussions often rely on the second sense, where phenomena in higher level sciences are reduced to the lower levels of biology. This sense is invoked by Sober's focus on reducing lung cancer to lower level molecular and biochemical details, discussions of reducing classical Mendelian genetics to molecular biology,<sup>18</sup> and Bickle's "ruthless reduction" (Bickle, 2006).<sup>19</sup> When this second sense of reduction is employed it is often conflated with the third more stringent sense. Although I do not have the space to argue for it here, I find the third position to be exceedingly difficult to support on the grounds of explanatory practice in biology and neuroscience. This is because very few explanations in these sciences cite details from the lower levels of physics. Attempts to support the presence or plausibility of this third sense of reductive explanation in these sciences have the baffling character of claiming to represent current explanatory practice, while invoking a future "completed physics" that assesses explanation "in principle" as opposed to explanation "in practice."<sup>20</sup> The first sense of reduction strikes me as a much weaker reductionist claim than is argued for in the relevant debates. I find the second sense to be the most charitable representation of many claims in support of reductive explanation in these debates. For these reasons I will adopt the second sense of reduction in

---

<sup>18</sup>(Schaffner, 1993; Waters, 1990; Kitcher, 1984).

<sup>19</sup>In all of these examples it is assumed that reductive explanation (or theory reduction) is achieved by reducing to the lowest levels of biology, which the third sense of reduction denies.

<sup>20</sup>(Sober, 1999, 543), (Waters, 1990).

my analysis.

#### 4.4.1 Causal heterogeneity and the final common pathway strategy

Consider three examples of causally heterogeneous higher level disease phenotypes: (1) retinitis pigmentosa, (2) Parkinson’s disease (PD) and (3) drug addiction. Retinitis pigmentosa (RP) is a disorder characterized by deterioration of photoreceptor cells in the retina that leads to a loss of peripheral vision and ultimately blindness. This disease is characterized by “immense genetic heterogeneity” as it can be caused by mutations in at least 12 different genes.<sup>21</sup> Parkinson’s disease is characterized by progressive neurological decline involving symptoms of dementia, resting tremor, rigidity, and postural instability. The heterogeneous causes of PD include: single gene variants, single environmental influences, and various combinations of these factors.<sup>22</sup> Finally, the shared behavioral phenotype of drug addiction can be caused by drugs that differ in their chemical composition, e.g. it can be caused by drugs including nicotine, alcohol, and opiates (Nestler, 2005).

In these situations scientists view the heterogeneous causes as problematic for attempts to understand and explain the higher level disease phenotypes at the population level. This is motivated by the widespread view that disease phenotypes are defined by some shared causal etiology at the level of the human population.<sup>23</sup> In the absence of a common causal process—either in the form of a higher level cause that is multiply realized by all the lower level heterogeneous causes (as in the smoking example) or any other lower level pathology—scientists suggest one of two approaches: continuing the search for some common etiology of the original phenotype or dividing it up on the basis of its known heterogeneous causes. Debate over which approach to take often involves arguing over whether the original disease phenotype is a single unified entity or, alternatively, a clinical concept that improperly lumps heterogeneous disorders. In the context of PD this has led to a situation where the “traditional view of Parkinson’s disease as a single clinical entity is under scrutiny.”<sup>24</sup> As

---

<sup>21</sup>(Kennan, Aherne, and Humphries, 2005, 108).

<sup>22</sup>(Shulman et al., 2011).

<sup>23</sup>In fact this is often used as a criterion to judge whether clinically accepted disorders are legitimate disease entities or not. This view is found in both the biomedical (Kendler, 2011) and philosophical discussions (Kincaid and Sullivan, 2014).

<sup>24</sup> (Obeso, Rodriguez-Oroz, Goetz, Marin, Kordower, Rodriguez, Hirsch, Farrer, Schapira, and Halliday,

Shulman et al. state:

Although PD was initially recognized and described as a purely clinical syndrome, recent progress has splintered the unitary conception of this disease into a number of alternative views. Some have suggested embracing everything that behaves clinically as PD under a single diagnostic umbrella. Others have argued in favor of abandoning PD as a single clinicopathologic entity, instead enumerating many subtypes on the basis of varying clinical features, familiarity, and autopsy findings. (Shulman et al., 2011, 214).

The failure to identify a “unifying mechanism” for the disorder and increased awareness of its heterogeneous causes has left scientists wondering “Is Parkinson’s disease a single disorder?” and “Is a unitary model for Parkinson’s disease possible?”<sup>25</sup> Similar concerns are raised in the cases of RP and drug addiction.<sup>26</sup> If these cases involve clear lower level heterogeneous causes, why would scientists resist the move to divide up the disease phenotype on the basis of these causes? One important reason for this is that although the lower level causes of the disease differ across cases in the population, the higher level disease phenotype does not differ in any identifiable or relevant way across these cases. In the case of PD, although different cases have different lower level causes, these cases are “clinically indistinguishable” at the level of the disease phenotype (Klein and Schlossmacher, 2006, 137). As these disease phenotypes are comprised of very unique groups of symptom clusters that reoccur in the population, these features seem to support treating the symptom-cluster as a shared higher level disease. Of course, scientists want to know what explains these symptom-clusters and why they are caused by different lower level factors. Dividing them up can prevent the pursuit of these questions, so the phenotype is often kept undivided while search for some common etiology continues.

One strategy scientists use to explain these disorders involves identifying a “common pathway” or a “final common pathway.”<sup>27</sup> This involves finding some shared causal process that all lower level (or upstream) heterogeneous factors converge on, and operate through, in producing the disease phenotype. The causal pathway allows for the identification of shared

---

2010, 653).

<sup>25</sup> (Obeso et al., 2010, 655, 653, 659).

<sup>26</sup>The debate is perhaps less contentious with drug addiction, as the default to divide up the phenotype on the basis of different chemical drugs is viewed as more acceptable in the biomedical community.

<sup>27</sup>For an informative discussion of the final common pathway concept see (Schaffner, 1998, 241), (Schaffner, 2008, 76). Aside from this work, the concept of a final common pathway has received relatively little attention in the philosophical literature.

causal factors for the type level effect of interest, where such factors meet the interventionist criteria for causation. More specifically, these pathways allow for the identification of factors that (when ideally intervened upon) can allow for control over the type level phenotype.<sup>28</sup> When such final common pathways are found, scientists appeal to them in explaining the type level disease and in identifying areas to target in treating all cases of a particular disease. For all three causally heterogeneous disease examples above, scientists claim to have identified a final common pathway that explains these diseases and reveals their distinct shared causal etiology. In the case of RP, “it is generally recognized that, regardless of differences in the underlying genetic cause, the ‘final common pathway of retinal degeneration’ is the apoptotic death of the photoreceptors.”<sup>29</sup> In PD this pathway involves the death of dopaminergic neurons and subsequent circuitry alterations that lead to the disease.<sup>30</sup> In the case of drug addiction, evidence suggests that “[d]rugs of abuse, despite diverse initial actions, produce some common effects” which can be thought of as a “common molecular pathway” leading to addiction (Nestler, 2005, 1445). As Doweiko states, “the important point is that all of the drugs of abuse (including alcohol) active the same nerve pathways involved in the process of learning/memory formation in addition to the reward circuitry in the brain (Coreia, 2005; Wolke, 2006). They share this final common pathway in spite of differences in their route of administration or chemical structure. From this perspective, the disorder of addiction might be viewed as one with multiple forms (activating chemicals) but a common etiology (Schaffer et al., 2004).”<sup>31</sup> As a final illustration of this strategy consider Cannon and Keller’s “Watershed” model of schizophrenia, in Figure 15, which reveals the convergence of lower level (upstream) gene variants on various common causal process that led to the disease phenotype (Cannon and Keller, 2006). Although different cases of Schizophrenia are caused by different gene variants (or combinations of them in addition to non-genetic factors), evidence suggests that that the higher level disorder may have a common causal etiology at the higher levels of neuronal development and cognitive processing.

---

<sup>28</sup>This is in contrast with other shared “causes” that do not meet the interventionist criterion of causation and do not allow for such control. These factors are often viewed as background conditions, like oxygen, which are required conditions for other causal relationships to operate.

<sup>29</sup> (Kennan et al., 2005, 108).

<sup>30</sup>(Burbulla and Krüger, 2011; Lesage and Brice, 2009; Corti et al., 2011).

<sup>31</sup> (Doweiko, 1999, 34).

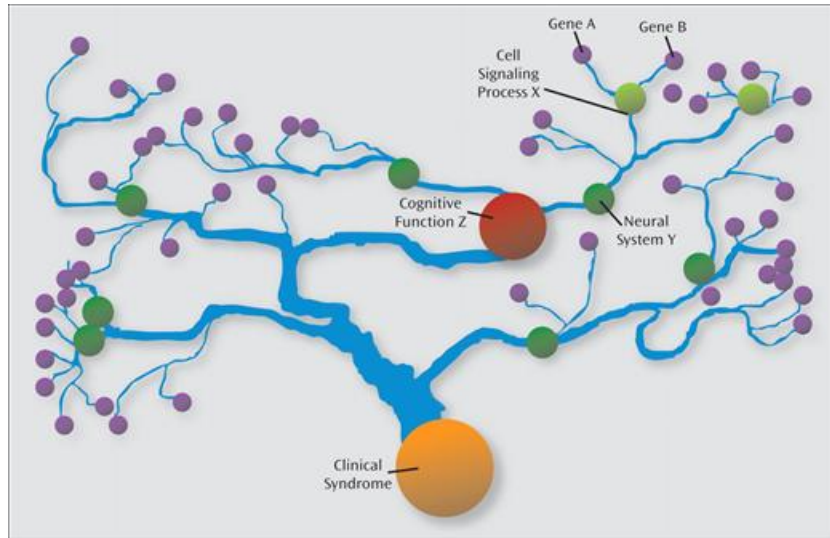


Figure 15: Convergence of gene variants on common processes: Cannon and Keller’s “Water-shed model of the pathway between upstream genes and downstream phenotypes” (Cannon and Keller (2006), p. 273)

In these situations, the location of the final common pathway dictates the type and level of causal details cited in explaining these disease phenotypes. Furthermore, the location of the final common pathway can be anywhere from the (1) lower levels of biology, (2) more intermediate levels, or the (3) higher levels of the disease phenotype.<sup>32</sup> For an example of the first situation, consider a case where heterogenous chemical compounds all interact with the same molecular receptor and produce the same downstream causal effects. In this case the shared causal etiology would start at the level of the chemical receptor, as it is the convergence point for the heterogenous causes. An example of the second situation includes genetically heterogeneous disease phenotypes like RP or PD, where lower level gene variants converge on processes at intermediate levels, like cellular or neural circuitry. While these common causal processes are at a higher level than gene variants, they are viewed as being at a lower level than the behavioral manifestations that characterize these disorders.

<sup>32</sup>In some cases it makes sense to view the shared causal etiology as spanning multiple levels, as in the case of PD where the common causal pathway spans from dopaminergic neuron functioning, to neural circuitry, to the disease. See (Kendler, 2011).

The third situation can represent various types of medical disorders resulting from physical trauma. For example, subdural hematoma and traumatic aortic rupture are both caused by high impact injuries, like those resulting from motor vehicle accidents and serious falls. The immediate or direct causes of these phenomena are high velocity impacts, which are viewed as higher-level causes (i.e. car crashes, falling, etc.), despite their being different identifiable lower level causally relevant factors that differ across impact scenarios (e.g. differing lower level features of the collision object).

These claims are consistent with and supported by other discussions of reductive explanation in the biological sciences. That these common causal processes can arise at any level is suggested by Schaffner who states “investigators need to be attentive to the possibility of common pathways emerging at any stage (early, intermediate, and final) in the temporal evolution of a reticulate network and involving multiple causes and complex “crosstalk” (Schaffner, 1998, 241). Other treatments of reduction that have elucidated the importance of “inter-level phenomena” and “theories of middle-range,” that acknowledge the important and frequent role that such intermediate level phenomena play in explanation in biology and neuroscience (Wimsatt, 2007; Schaffner, 1993). This supports similar views on reduction that genuinely acknowledge the relevance of both lower level and higher level details, without privileging one over the other. As Wimsatt writes, “[n]either extreme is acceptable: both fail to capture the integrative articulation in a single mechanistic explanation of entities, phenomena, and causes from different levels—the basic form of explanations that are so common in modern scientific and engineering practice” (Wimsatt, 2007, 173).

What does this reveal about the potential for reductive explanation in these situations? I have indicated that scientists often explain these causally heterogeneous type level phenomena by appealing to some common causal etiology, sometimes called a common pathway. In these cases, the location of the final common pathway dictates the type and level of causal details cited in explaining these disease phenotypes, just as the location of the common cause “smoking” dictated the level of detail cited in explaining the type level phenomenon of lung cancer. However, as the common causal etiology for these phenomena can be located at any level from lower, intermediate, to higher levels—there is no single level of causal detail that all such explanations will appeal to. That being said, there will be cases where such

causal heterogeneity prevents reductive explanation, when the common causal process of some type-level effect is found at intermediate or higher biological levels. There will also be situations where reductive explanation is perfectly compatible with such cases of causal heterogeneity, when both the heterogeneous causal factors and the shared etiology are located at the lower levels of biology. These will be situations where the causal complexity of lower level details does not lead to non-reductive explanations or the autonomy of the higher levels. Finally, neither the reductive or non-reductive position will account for explanations that cite intermediate causal processes and details, which seem to make up many (if not most) explanations of these phenomena. There are no a priori limitations on which level of causal detail is relevant to all causal explanations in these sciences. There are explanatory limitations in citing heterogeneous lower level causes for type level phenomena, but these situations are compatible with both reductive and non-reductive explanation.

In these cases, whether an explanation of some type level phenomenon is reductive or not depends on the level at which the common causal etiology is located. The location of this causal process depends on the explanatory target of interest and empirical features of the world. Looking at the watershed model in figure 14, we can consider an explanation of a token instance of the clinical syndrome in an individual, for which gene A is causally responsible for the disease. In this case there will be fewer restrictions on appealing to lower level details unique to this patient, compared to a population level explanation of the disease. Changing the explanatory target to some type phenomenon in a heterogeneous population, where heterogeneous causes are implicated, constrains which causes are shared in the population and thus, which causes can be cited in this explanation. In both cases, empirical features of the target of interest matter, and while they change depending on the explanatory target they are still “fixed” by actual features of the world. This makes sense of why the medical community focuses on intermediate and higher level causes in explaining, preventing, and treating the disease at the population level. Of course lower level details figure into explanatory and treatment practices when the focus is on an individual patient and this makes sense given that this involves a different explanatory target where the focus is on explaining and treating the effect in one individual. Biology and neuroscience have room for various types of explanatory questions and explanatory targets. The causal details



that matter for such explanations are determined on the basis of these questions and targets, and empirical features of the world.

## 4.5 CONCLUSION

I have argued that a particular type of multiple realizability can prevent reductive explanation in biology and neuroscience, but that this does not imply the strong form of autonomy that some supporters of this thesis advocate. I indicated how this form of multiple realizability relates to cases of causal heterogeneity, which pose similar problems for reductive explanation. These problems are easily overlooked if various differences between type and token level explanations are not considered. Such considerations are important, because explanations in biology and neuroscience focus on explanatory targets that include both type and token level phenomena. Physicians treating individual patients will focus on explaining and understanding token level instances of disease. Public health professionals and epidemiologists treating large communities and patient populations will focus on type level instances of disease. Focus on explaining disease at both targets makes sense given our goals of treating disease in both individual patients and patient populations. As multiple realizability and causal heterogeneity involve differences across systems or individuals, these features will have influences on type level explanation that may not be noticeable with a focus on token level phenomena. However, the cases of causal heterogeneity discussed in this chapter indicate that biology and neuroscience have room for both reductive and non-reductive explanations. The type of explanation provided depends on the explanatory target of interest and empirical features of the world.

## 5.0 CAUSAL CONTROL: A RATIONALE FOR CAUSAL SELECTION

### 5.1 INTRODUCTION

Causal selection has to do with the distinction we make between background conditions and “the” true cause or causes of some outcome of interest. We claim that “the” cause of a match lighting was the fact that it was struck, but not the presence of oxygen. A car crash may involve many causal factors—ice on the road, the speed of the car, the gas in the tank, and so on—but we are likely to explain it by citing few of these. In each case, just a small number of factors are selected as causes of the outcome, while most others are backgrounded. A longstanding consensus in philosophy views causal selection as lacking any objective rationale and as guided, instead, by arbitrary, pragmatic, and non-scientific considerations. This position is famously supported by John Stuart Mill, whose arguments have “won the field” and remain “echoed by contemporary authors” ([Schaffer, 2014](#)). According to Mill, “Nothing can better show the absence of any scientific ground for the distinction between the cause of a phenomenon and its conditions, than the capricious manner in which we select from among the conditions that which we choose to denominate the cause” ([Mill, 1874](#), 238). Lewis (1973) agrees with this, claiming that we select causes because they are under our control, because we find them good or bad, or just because we want to talk about them. He states, “We sometimes single out one among all the causes of some event and call it ‘the’ cause, as if there were no others. Or we single out a few as the ‘causes,’ calling the rest mere ‘causal factors’ or ‘causal conditions’... I have nothing to say about these principles of invidious discrimination” ([Lewis, 1973](#), 558-559).

This position faces significant problems in the context of biomedicine, where scientists commonly identify “the” cause or causes of specific diseases. Scientists claim that the cause

of scurvy is vitamin C deficiency, that the cause of tuberculosis is the tubercle bacterium, and that the cause of Huntington’s disease is a mutation in the *huntingtin* gene. There is widespread consensus on the selection of these causes in the medical community. Furthermore, at first glance, these causes appear importantly relevant to their respective diseases in ways linked to possibilities of control. Targeting these causal factors has allowed for successful prevention and treatment of disease, to the extent of drastically reducing the incidence of some diseases and nearly eradicating others from the human population. If causal selection is arbitrary, why is there such widespread consensus on the selection of causes for some disease traits? What explains the apparent success of this selection in identifying factors that reliably allow for control over disease? Finally, if causal selection is unscientific, why do we view disease explanation as a scientific matter? These points identify significant problems for the mainstream philosophical position on causal selection and its ability to account for the selection of disease causes in biomedicine.

In this chapter I argue that the rationale behind causal selection for disease is best understood in terms of the *causal control* that selected causes have over the disease of interest. I provide a novel account of the types of control that guide this process. In this domain, causes are selected on the basis of having (i) specific, (ii) probable, and (iii) stable control over disease. I suggest that this selection is pragmatic in the sense that it is relative to a *practical* goal—the goal of control. This goal is practical in the sense that it is useful, in general, for our navigation and operation within the world and, more specifically, for our practical aims of treating and preventing disease in patients. This is a different notion of “pragmatic” than is often found in the philosophical literature, where it is used to refer to considerations that are arbitrary, subjective and/or audience-relative ([Achinstein, 1984](#); [Schaffer, 2014](#)). Furthermore, I suggest that causal selection for disease is also “objective” in the sense that once the goal of control is specified, there are objective facts and considerations about what means conduce to this goal. This chapter provides an analysis of causal selection that clarifies these considerations.

I characterize causal selection as involving three main steps. The first step involves setting a contrastive focus, while the second and third steps involve selecting causes with respect to this focus. In order for a factor to be selected as “the” cause or one of “the” causes

of an effect, it must pass both the second and third steps. Once the (1) contrastive focus is set, the remaining two steps involve selecting causes on the basis of the causal control they have over this focus. The second step involves (2) assessing whether a factor has *any* control over the effect of interest, and the third step involves (3) assessing what *type* of control the factor has over the effect. I clarify three types of control that guide this third and final selection step, which include control that is (i) specific, (ii) probable, and (iii) stable for the disease of interest. It makes sense that we select causes on these grounds, because they meet particular standards we have for disease explanation and they provide valuable targets for disease treatment and prevention. An important feature of my analysis is that it clarifies how unique features of disease traits make causal selection for disease much different, and arguably much easier, than causal selection for many other types of phenomena. One reason for this, is that scientists often impose constraints on what qualifies as a “legitimate” or “valid” disease category, where these constraints alone significantly narrow the number of candidate causes for any given disease. Differences between disease examples and the examples commonly examined in the philosophical literature on causal selection can help clarify why mainstream philosophical views have overlooked the principled rationale that guides causal selection for disease.

This chapter is organized as follows. Section 2 discusses important features of disease traits and how they set the relevant contrastive focus. With this contrastive focus specified, I provide a minimal criterion for causal control that captures the second step of causal selection. This criterion draws on Woodward’s (2003) interventionist account of causation and it identifies “interventionist causes” for the contrastive focus. Section 3 examines the third and final selection step, where particular factors are selected from a pool of candidate causes. I clarify the types of control that guide the selection of causes in this final step. Throughout this analysis I clarify how causal selection for disease is influenced by both pragmatic and objective considerations and I return to this topic in the the concluding section.

## 5.2 DISEASE TRAITS AND INTERVENTIONIST CAUSES

Human diseases are a unique type of biological phenomenon. Modern medicine aims to define disease traits causally, in terms of particular causal etiologies. While there is a sense in which all diseases are produced by a multitude of causal factors, often very few factors are selected as “the” cause or one of “the” causes of any given disease. For “monocausal” diseases we cite single factors, as in the cases of scurvy, tuberculosis, and Huntington’s disease. For other diseases we cite longer lists of causes. For example phenylketouria (PKU) is explained by appealing to both a gene mutation and a dietary factor.<sup>1</sup> When little is known about the causal etiology of a clinically accepted disease category—as is common with psychiatric and other disorders—the “legitimacy” and “validity” of the disease is often disputed or, at least, considered an open question.<sup>2</sup> In this chapter, I focus on diseases for which we have some sufficiently understood causal etiology. These are the cases where we commonly select disease causes and where there is consensus on this selection.<sup>3</sup>

Human diseases, even those with well understood or simple causal etiologies, often involve many symptoms. For example, although tuberculosis is explained by appealing to a single bacterial factor, patients with this disease often present with a wide variety of symptoms. These patients can exhibit symptoms that include: dry cough, blood-tinged sputum, night sweats, weight loss, and fatigue, to name a few. Although we distinguish among these symptoms we do not view them as individual diseases, but as features of a single disease process. Additionally, symptomatic presentation can vary significantly across patients with the same disease. Patients with the same disease may present with completely different combinations of symptoms or with similar symptoms that vary in degree. For an example of the former, one patient with tuberculosis may exhibit of all the symptoms mentioned above,

---

<sup>1</sup>PKU is characterized by disordered metabolism of the amino acid phenylalanine, which results in impaired neurological development. This disease is caused by both the ingestion of phenylalanine and a mutation in the gene for the phenylalanine hydroxylase enzyme.

<sup>2</sup>The worry is not that the symptoms and signs of the disorder are not real (in the sense of being experienced or exhibited by the patient), but that the disorder category will change with further clarification of the etiology (Kincaid and Sullivan, 2014). For more on this see Schaffner’s discussion of validity and etiopathological validity in the context of psychiatry and general biomedicine (Schaffner, 2012).

<sup>3</sup>If we know little about the causal etiology of a clinically accepted disease category, we typically do not consider ourselves in a position to identify its causes.

while another may present with only a dry cough. In the case of the latter, two patients with tuberculosis may present with a dry cough, but the severity of their cough may differ.

Tuberculosis has a relatively simple causal etiology in the sense that it is explained by appealing to a single causal factor. It also has a specific causal etiology, in the sense that all cases of this disease have the same cause. What this reveals is that a disease can have a simple and specific causal etiology, *without* this implying that it always presents in patients with a single, uniform symptomatology. Unfortunately, the view that particular diseases manifest in patients with uniform symptomatology is common in the philosophical literature.<sup>4</sup> This view incorrectly posits uniform symptomatology as a kind of “standard” for medically accepted disease categories. If distinct diseases always presented in such a uniform manner, diagnosis in the clinical setting would be much easier and straightforward than it is. Instead, patients with the same disease often present with highly divergent symptom profiles and clinicians are trained to diagnose in light of this challenge.

While we do not expect most diseases to have *simple* causal etiologies, we often assume that they have *specific* causal etiologies. In this sense, to say that disease D has a specific causal etiology means that all instances of D are produced by roughly the same causal factors.<sup>5</sup> (I discuss this further in section 3). As the notion of specific causal etiology pertains to many instances of a given disease—as opposed to a single or token instance—it represents a type-level consideration. This type-level focus is present in many of our claims about what “the” cause of a particular disease is. In this chapter, I focus on causal selection for type-level disease traits. This can be thought of as causal selection that focuses on answering the following question: “What is the cause of disease D in the human population?” When we answer this question we often focus on the binary contrast of disease “absence” and “presence.” Of course, a patient either has disease D or she does not—she cannot be in both of these states or in neither of them. We expect disease causes to explain *at least* this contrast, in part because it is often the ultimate contrast we want control over. If a causal factor only

---

<sup>4</sup> For examples of this, see: (Poland, 2014; Blaxter, 2015; Kincaid and Sullivan, 2014; Murphy, 2014).

<sup>5</sup>This is *not* the same as claiming that the causal factors in question *only* produce disease D and not other diseases—this refers to the specificity of *effects*, given some cause. The notion of specific causal etiology I discuss involves the specificity of *causes* given some effect (a particular disease). These two types of specificity may be distinguished as (2.1) specificity of effect (given some cause) and (2.2) specificity of cause (given some effect).

explains positive degrees of disease pathology, without also explaining disease absence, we view it as an incomplete and unsatisfying explanation. Our interest in completely curing and preventing disease—i.e. ensuring the complete absence of disease—motivates our interest in identifying causes of this binary contrast. If these causes also account for varying states of a disease—in addition to disease absence—we view this as an added advantage, but not a necessary criterion for the factors we select as disease causes.

The first step of causal selection involves setting a contrastive focus. So far, I have identified two important features of disease traits that set this focus: disease traits are (1) type level phenomena, which (2) are often represented as taking on the values “present” or “absent.” In this chapter, I argue that the rationale behind causal selection for disease is best understood in terms of the *causal control* of factors over this contrastive focus. In the second step of causal selection, a factor is selected as a candidate cause if it has *some* control, or some minimal amount of control, over this contrast. What does it mean for a factor to have *some* causal control? Consider a minimal interventionist criterion for causal control, which is met by factors I call “interventionist causes” and inspired by Woodward (2003):

**(i.c.) interventionist cause:** a factor C has causal control over disease D if and only if there are circumstances S such that if some (single) intervention that changes the value of C (and no other variable) were to occur in S, then the value of D or the probability distribution of D would change, *for the contrastive focus in question*.

The notion of an intervention ensures that when variable C is manipulated, it allows for a change in the value of D in a way that excludes confounders or other variables that may causally influence D. One advantage of the interventionist framework is its ability to capture the motivation behind some of our experimental methods for identifying causal relationships. If we want to determine whether substance X causes disease Y, we might design an experiment in a model organism where we manipulate values of X (and only X) to see if this causes a change in the occurrence of disease Y. In this experiment, we are likely to keep potential confounding factors constant (e.g. diet, exercise, etc.) in order to ensure that

changes in X—and not changes in these other factors—cause the outcome we measure.<sup>6</sup> The interventionist criterion (i.c.) involves a counterfactual claim: it maintains that C has causal control over D in the sense that *if* there was a change in C, this *would* produce a change in D. This criterion does not require that such an intervention on C is actually performed, in the sense of being manipulated or even manipulable with current technology or by human means. This makes sense of the fact that we view gene variants as the cause of many human diseases, despite our inability to intervene on them in human patients. Technological limitations and ethical restraint prevent us from manipulating these gene variants in humans. However, we still maintain that some genes cause particular diseases, in the counterfactual sense indicated: we mean that *if* such gene variants were manipulated, this *would* change the disease status of the subject.<sup>7</sup> We support this claim on the basis of evidence acquired from various sources and not just from actually performing the intervention in question. My analysis relies on a notion of causal control that is counterfactual in the same sense.

In order for a factor to be selected as a disease cause, it must meet (i.c.) and have *some* causal control over the contrastive focus. This is the second step and the first “cut” of the causal selection process. Notice that the causes we select for scurvy, tuberculosis, and Huntington’s disease all meet this interventionist criterion (i.c.). Dietary vitamin C has causal control over scurvy in the sense that manipulating levels of this dietary factor provides control over the occurrence and nonoccurrence of the disease.

It might seem that the (i.c.) criterion is overly inclusive, in the sense that it picks out a huge number of factors. Recall the car crash example, for which numerous factors were identified as causally relevant. Lewis claims that the causes of the crash include: the ice on the road, the blind corner, and even the birth of the driver’s paternal grandmother (Lewis, 1986, 215-6). If all these factors meet the (i.c.) criterion, it might not seem like a very helpful first

---

<sup>6</sup>For more of these details, see (Woodward, 2003).

<sup>7</sup>This highlights important differences between my position and Gannet’s (1999) analysis of pragmatic considerations that influence genetic explanation. She views causal selection as guided by our success with *actual* manipulation of candidate causes. I think her position neglects the importance of counterfactual (or hypothetical) information in causal selection. Acknowledging the importance of this information helps explain why we cite genes and other factors as disease causes, despite our limitations in actually manipulating them. Furthermore, Gannet and I both claim that causal selection in biology is pragmatic, but we disagree about why. I suggest that causal selection is pragmatic in the sense that it relates to the practical goal of control, as opposed to our ability to actually manipulate causal factors.



cut. The ordinary life examples discussed in the philosophical literature are different from disease examples, in a way that can misrepresent causal selection in biomedicine. Disease traits are very narrow phenomena, both by our choosing and for reasons that have to do with their manifestation in living organisms. First, as we often define disease traits in terms of specific causal etiologies, our own characterization of them significantly constrains their relevant causal factors from the outset. (I discuss this more in section 3). Second, disease phenotypes are narrow in the sense that their causes are expected to control the disease absence/presence contrast, and only this contrast, in living human patients. This is a very narrow and fragile contrast for a factor to have control over, in a way that is not characteristic of many non-biological examples. If we want to prevent a match from lighting, or put it out once lit, there are many ways to do this. We could chop the match into unrecognizable pieces, pour corrosive chemicals on it, or throw it in the ocean. There are far fewer candidate causes for disease, because we expect such factors to have control over disease traits without killing or harming the patient. We can destroy the match to prevent it from lighting, but we do not want to destroy the patient to eliminate disease. There is a sense in which there is no disease in a dead patient, but clearly this is not the type of control that we want, or expect, disease causes to have. These features of disease traits significantly reduce the number of candidate causes that we consider in our search for “the” cause or causes of a given disease.

Once the contrastive focus of disease presence/absence is specified, the (i.c.) criterion clarifies “objective” considerations that guide selection. Given the goal of control, there are objective facts about which factors have control over this focus and which factors do not. Consider the disease scurvy again, but now with regard to the candidate causes “oxygen” and “dietary vitamin C.” “Oxygen” does not meet the (i.c.) criterion for this disease, because manipulating oxygen does not allow for control over the “absence” and “presence” of scurvy in *living* human patients. Of course, manipulating oxygen in certain ways can kill a patient—this would happen if oxygen levels were set to a very low value or if oxygen were completely removed from the patient’s environment. Manipulating “oxygen” does have some control over whether a patient lives or dies, but this is *not* the dominant contrastive focus for disease explanation. Instead, the dominant focus is disease absence/presence in *living* patients. “Dietary vitamin C” meets the (i.c.) criterion for this contrast in the case

of scurvy, while “oxygen” does not. This explains why we cite dietary vitamin C, and not oxygen, as “the” cause of scurvy. However, while “oxygen” does not meet the (i.c.) criterion for scurvy there is still a sense in which it is relevant or “necessary” for the incidence of this disease. “Oxygen” is relevant to this disease in the sense that its presence is required for the causal control that vitamin C levels have over scurvy. This is due to the fact that oxygen is a requirement for human life and thus, it is required for attaining both values of the contrast in question. This clarifies how we can view a factor like oxygen as importantly relevant or necessary for disease, while still distinguishing it from factors we select as “the” cause or causes of disease.

Thus, if a factor lacks causal control over the contrastive focus for a disease trait it fails to pass the second causal selection step. This reveals part of the rationale behind causal selection for disease traits and how the medical community can reach consensus on “the” causes of some diseases. This consensus is partly explained by our expectation that disease causes should have some causal control over the disease of interest. The interventionist criterion (i.c.) captures this basic standard and the second step of the causal selection process. In the next section I discuss the third step of this process, where we select among candidate interventionist causes for a given disease.

### 5.3 SELECTING AMONG INTERVENTIONIST CAUSES

In order for a factor to be selected as “the” cause or one of “the” causes of a disease, meeting the interventionist criterion (i.c.) is necessary, but not sufficient. Some factors meet (i.c.) for particular diseases without being selected as “the” cause of the disease. For example, consider sleep deprivation and the seasonal flu. Changing sleep duration, such that a patient is sleep deprived, has some causal control over the flu in the sense that it can increase susceptibility to infection by decreasing immune functioning (Bryant, Trinder, and Curtis, 2004). However, we do not consider sleep deprivation “the” cause (or even one of “the” causes) of the seasonal flu. We reserve this designation for the particular flu virus. What explains this selection of one interventionist cause over another? Notice that targeting the flu virus (e.g. with

vaccination<sup>8</sup>) provides a very different type of control over the occurrence of the flu than does targeting patients’ amounts of sleep. Targeting the virus provides (3.1) control over many to all cases of this particular flu, (3.2) a high likelihood of preventing it in each case, and (3.3) control across a wide variety of genetic and environmental conditions present in the patient population. Compared to the flu virus, targeting levels of sleep significantly underperforms in all of these areas. I discuss these three types of causal control—which I refer to as causal control that is (3.1) specific, (3.2) probable, and (3.3) stable—and I clarify how they guide our selection of disease causes.

### 5.3.1 Specific causes and causal control of broad scope

The factors we select as disease causes exhibit very particular types of causal control over disease. The first type of causal control I discuss relates to the assumption of specific causal etiology. This assumption captures our default view that type-level disease traits have specific causes. In other words, for a given disease D we often expect that most or all instances of D are the result of a similar causal process. In the philosophical literature, the notion of specific causal etiology for disease is often referred to as the “causal signature” (Murphy, 2014, 105), “disorder-specific pathophysiology” (Caspi and Moffitt, 2006, 586), or “shared causal process” (Zachar, 2014, 87) for a disease trait. Sometimes we refer to single factors as the specific causal etiology for a disease, like the *huntingtin* gene mutation for Huntington’s disease. Other times we refer to multiple interacting factors as the specific causal etiology for a disease, like the gene variant and dietary factor for the disease PKU. The important feature of specific causal etiology is not how many factors cause an instance of disease D, but that the *same* factors produce all or most instances of disease D.<sup>9</sup>

Just because we expect diseases to have specific causal etiologies, does not mean that all clinically useful disease categories meet this standard. Sometimes we start with clinically useful categories—like Parkinson’s disease—only to find out later, that these categories fail to

---

<sup>8</sup>This is the vaccine that is recommended annually, for everyone 6 months and older.

<sup>9</sup>In this chapter, I use a notion of specificity that refers to the number of cause and effect *variables* that participate in a type-level causal relationship, as opposed to the *values* of these variables. The latter (*value*) sense of specificity has received much more attention in philosophy of biology (Woodward, 2010; Waters, 2007; Griffiths, ). For more on this distinction, see (Woodward, , Forthcoming).

track specific causal processes. (This is unsurprising, because we often create these categories before we have clear information about their causes.) For example, our best evidence suggests that Parkinson’s disease has a *non-specific* causal etiology, in the sense that (3.1.1) different combinations of causal factors cause instances of the disease on different occasions. This contrasts with a situation of *specific* causal etiology where (3.1.2) several factors interact to produce a disease, but where every case of the disease is produced by the same interacting factors. An example of (3.1.2) is PKU, because the same two interacting factors cause every instance of the disease.

What does the assumption of specific causal etiology have to do with selecting disease causes and causal control? This assumption clarifies our aim of selecting factors that are specific for a disease, in part because these factors are likely to provide control of *broad* scope over all or most instances of the disease in question. Consider the situations above, where a disease D has either (3.1.1) a non-specific causal etiology or (3.1.2) a specific causal etiology. One advantage of selecting specific causes is that they can often be manipulated to control all or many cases of the population-wide disease of interest. This is because, in situation (3.1.2), all cases of the disease are produced by the same causal factors. Alternatively, if we select causes that are non-specific for a disease, as in the case of (3.1.1), these factors are likely to have causal control of narrow scope, in the sense that they influence a smaller percentage of the total cases of disease D. This is because, in situation (3.1.1), different causal factors produce different instances of the same disease D, so targeting the factors that produce any one instance of the disease, is unlikely to provide control over the other instances that have different causes. From the standpoint of treating and preventing disease, identifying factors that are causally specific for particular disease traits is extremely valuable. It often identifies factors that we can target to control and explain a large percentage of all cases of a given disease in the population.

The notion of specific causal etiology plays a more complicated role in causal selection than I have indicated so far. One complication is that this notion influences both the factors we select as disease causes *and* those phenomena that we consider to be “legitimate” diseases to begin with. If a disease category has no known causal etiology, or if it has a non-specific causal etiology, the “legitimacy” of the category and whether it represents a

“true” disease are often called into question. This is helpfully illustrated with Parkinson’s disease, which we understand as having a non-specific causal etiology. As mentioned above, current research suggests that different combinations of causal factors produce different cases of Parkinson’s disease. Researchers view this causal non-specificity as having “splintered the unitary conception of this disease” and as “challenging traditional conceptual frameworks” for understanding Parkinson’s disease (Shulman et al., 2011, 214, 193). In this situation medical researchers suggest either (1) continuing the search for some shared causal process (i.e. some specific causal etiology) or (2) dividing up the disease category on the basis of the distinct causal processes. Both of these options restore causal specificity by either (1) finding it for the pre-established disease category or (2) creating it by redefining the disease category. This reveals how causal specificity is both a guiding rationale for causal selection and a standard that influences how we define disease traits. This suggests that causal selection is more of a back-and-forth process than just a search for causes given a fixed contrastive focus. We may start with a disease category and search for its causes, only to redefine the category on the basis of what we find.

### 5.3.2 Probable causal control

A second type of causal control that guides causal selection for disease is what I call *probable* causal control. This causal control refers to the *probability* with which each outcome of the contrastive focus is produced when selected factors are manipulated. Consider a light switch C, which can take the values ‘up’ or ‘down’ and a light E, which can take the values “on” or “off.” In the first case, turning the switch “up” results in a 99% probability of the light bulb being “on” and turning the switch “down” results in a 99% probability of the light bulb being “off.” In a second scenario, turning the switch “up” has only a 60% probability of causing the light to turn “on” and turning the light switch “down” only has a 60% probability of turning the light “off.” In both cases the switches have some causal control over the state of the light, but their control differs with regard to how *probable* each outcome of the contrast is with interventions on the switch.<sup>10</sup>

---

<sup>10</sup>This notion of probable causal control shares similarities with the suggestion by Lu et al. that we look for causes with high “power” (Lu, Yuille, Liljeholm, Cheng, and Holyoak, 2008).

When we select disease causes we prioritize factors that have a high degree of probable causal control over disease. By targeting these factors, as opposed to factors with less probable causal control, we increase the likelihood of getting a particular outcome. This has a clear advantage for our treatment and prevention measures and for explaining disease outcomes. If we want to control whether a light is “on” or “off” we will prefer the first switch-bulb system to the second, because we are more likely to get the outcome we want by manipulating the switch. In cases where we reach consensus on disease causes, they often have probable causal control over the disease trait in question. Consider diseases like scurvy, tuberculosis, and Huntington’s disease. In each of these cases, when the disease cause is present (or properly introduced) in a patient, her likelihood of acquiring the disease approaches 100%. Similarly, when the cause is absent (or properly avoided) the likelihood of disease absence also approaches 100%. A patient with the *huntingtin* gene mutation is almost certain to acquire Huntington’s disease and very unlikely to get this disease without it (cases of this disease without the mutation are unheard of). Alternatively, consider our attempts at causal selection for schizophrenia, which is a psychiatric disorder associated with a multitude of causally relevant gene variants. In this case, no single variant (or set of variants) confers a high probability of disease occurrence and nonoccurrence, although some provide a low probability of this sort. Failure to meet the standard of probable causal control partly explains why the medical community does not view such variants as “the” cause of this disease and why the search to better understand its etiology continues.

When a single causal factor provides a low degree of probable causal control over a disease trait, we often search for interacting causes that increase this type of control. Consider PKU again, which we explain by citing two causal factors: a gene variant and a dietary factor. These are interacting causes for this disease, because they both meet the (i.c.) criterion and they each influence the causal control that the other has over the disease. The gene variant only causes PKU when the dietary factor is present, and vice versa. One reason for selecting both of these causes in explaining PKU is that together they provide more probable causal control over the disease than a single factor alone. Of course, PKU is a relatively simple disease in the sense that we explain it by citing only two interacting causes. One significant challenge associated with disease explanation is that many diseases appear to be causally

complex in the sense that they have a multitude of interacting causes, which must all be accounted for to provide a high degree of probable causal control over the disease. This can make disease explanation and causal selection much more difficult, because a larger number of causal factors have to be identified.

To say that causes with probable causal control are privileged in causal selection is different from claiming that causation requires this type of control. Some probability-raising accounts of causation support this later position, by maintaining that causes are factors that result in a high probability of their effects.<sup>11</sup> A well known objection to these accounts cites the low probability of general paresis, or late state neurosyphilis, among untreated patients with syphilis (Scriven, 1959). Approximately one-third of patients with untreated syphilis end up with general paresis, yet we still view the syphilis bacterium as “the” cause of this low probability outcome. Probability-raising accounts struggle to make sense of why we explain general paresis by citing the syphilis bacterium, since the cause confers a low probability of the occurrence of the effect, compared to a causal factor that approaches a 100% likelihood of producing an outcome. My analysis clarifies this confusion and explains why this example is not problematic for my position. First, we more often view general paresis as a set of *symptoms* produced by the disease syphilis, as opposed to a distinct *disease* itself. It is true that we view the syphilis bacterium as the cause of general paresis, but we rarely expect disease causes to have specific or probable causal control over disease *symptoms*. This is because individual symptoms can be found in many different diseases (and thus, have many different causes) and they can have variable presentation across cases of a particular disease (where there is, presumably, the same causal etiology).<sup>12</sup> When we focus on the proper disease target “syphilis,” we do identify the syphilis bacterium as “the” cause of this disease. Our reasons for doing this include the facts that this bacterial cause exhibits specific and probable causal control over the disease trait. Furthermore, my analysis clarifies what features of causal relationships we privilege in some contexts, but not what features of relationships make them causal. To say that factors with less probable causal control are not privileged in causal selection is not to deny that they are still causal relationships.

---

<sup>11</sup>Hempel’s inductive statistical (IS) account of explanation has been viewed as supporting this position (Hempel, 1965).

<sup>12</sup>I discuss these points in section 2.

### 5.3.3 Stable causal control

A final type of causal control I discuss is causal control that is stable. Stability is a feature of causal relationships that has been discussed extensively by Woodward (2003, 2006, 2010). Recall that the interventionist criterion (i.c.) refers to “circumstances  $S$ ” in which a cause  $C$  has causal control over an effect  $D$ . Stability refers to the extent to which the causal control of  $C$  over  $D$  holds in a range of other circumstances  $S_i$ , which differ from circumstances  $S$  (Woodward, 2010). Consider a broad set of conditions, which include the various genetic backgrounds of the current human population and the range of environmental surroundings they live in. If a cause  $C$  only has causal control over effect  $D$  in a very narrow range of these circumstances, this causal control is *unstable*. An example of this would be a situation where manipulating dietary vitamin C levels only controlled scurvy incidence in people with a gene variant for brown hair, who also live in the state of Florida.<sup>13</sup> In this case, the causal control of  $C$  over  $D$  only holds in patients with a narrow range of genetic and environmental conditions, relative to all patients and environments in the world. Alternatively, if  $C$  has causal control over  $D$  in a very broad range of these circumstances, this causal control is considered *stable*. An example of this would be if manipulating dietary vitamin C levels controlled scurvy incidence in patients of a wide variety of genetic backgrounds, who live in many different environments. The second situation of stable causal control is clearly more useful for the purposes of treatment and prevention of disease—it allows for measures that have the potential to treat a larger number of patients who live in many different types of environments. The causes we select for scurvy and tuberculosis have stable causal control in the sense captured in the second scenario. We can target these causes to treat and prevent these diseases in patients with diverse genetic backgrounds, who live in a wide variety of environments.

There is an important complication involved in assessing the stability of causal relationships. The degree to which a cause variable has stable (or unstable) causal control depends on the range of circumstances  $S_i$  that are considered. Clarifying exactly how broad or narrow these circumstances are is difficult, because this determination often seems context-

---

<sup>13</sup>This is not a true scenario. It is intended to clarify the relevant sense of unstable causal control.



dependent. This is acknowledged by Woodward, who states that the range of circumstances that matter for assessing stability are those that “do not depart too much from the actual state of affairs or that do not seem too far-fetched or that are not judged to be unimportant or irrelevant for subject-matter-specific reasons” (Woodward, 2006, 11).

What are the circumstances  $S_i$  that matter for determining stability in the context of disease traits? First, there is an important sense in which stability is relative to particular reference classes, which group patients on the basis of factors like age group and sex, depending on the disease of interest.<sup>14</sup> For example, given a virus that causes cervical cancer, we do not assess the stability of this cause with respect to all patients, but only with respect to patients of the female sex.<sup>15</sup> The same can be said for pediatric, geriatric, and other diseases, which are present in particular patient populations.<sup>16</sup> When diseases are restricted to particular populations, we often assess the stability of causal relationships with respect to this restriction. Second, as Woodward has indicated, the circumstances that matter for assessing stability are those circumstances that *actually* occur in the contexts we are interested in. In the context of disease, we care about the range of conditions that are *actually* present in human patients and their environments in the world. We care about these circumstances because they are the circumstances in which we want to control disease. Whether a factor meets this standard depends on the time-frame of interest. Future changes in our genetic make-up and the environments we live in can alter the circumstances  $S_i$  that we use in assessing the stability of causal factors for disease. Third, we restrict the conditions included in circumstances  $S_i$  to those that include basic biological requirements for human life. Just because there are some low oxygen environments on Earth, does not mean that we assess the stability of disease causes with such circumstances in mind. All disease causes break down in this type of environment, because it is incompatible with human life. Since we cannot sustain human life in these circumstances, we are not focused on controlling disease in them.

In cases where we identify “the” cause or causes of disease traits, these factors often have control that is specific, probable, and stable, for the disease of interest. It makes sense

---

<sup>14</sup>Boorse provides a helpful basic characterization of a reference class as “an age group of a sex of a species” (Boorse, 1977, 555).

<sup>15</sup>The cervix is an organ that is present in females, but absent in males.

<sup>16</sup>Providing an account of reference classes and their role in stability assessments is outside the scope of this chapter.

that we privilege these factors, because they provide types of control that serve our interests in explaining, treating, and preventing disease. For a given disease trait, targeting such factors is likely to provide (3.1) control over many to all instances of the disease, (3.2) a high likelihood of preventing each instance, and (3.3) control across a wide variety of genetic and environmental conditions that are present in the patient population.

## 5.4 CONCLUSION

What does this analysis suggest about the role of “pragmatics” in causal selection for disease traits? In the philosophical literature “pragmatic” is commonly used to imply that something is arbitrary, subjective, or audience-relative. Causal selection for disease does not appear to be “pragmatic” in any of these senses. The significant consensus on causal selection for some disease traits and my analysis of the principled rationale that guides this selection suggests that it is not simply an arbitrary procedure. The factors we select as disease causes provide special types of causal control that an arbitrary selection method would not explain. Furthermore, it isn’t clear that subjective or audience-relative preferences, whatever they may be, could capture the sense in which causal selection for disease is relative to a practical goal, viz. the goal of control. Information relevant to control can provide us with the means to change disease outcomes of real patients. It isn’t just a contrived way of selecting causes that we want to talk about, that strike our fancy, or that we just happen to view as important, without good reason. The factors we select as “the” cause or causes of particular diseases often provide viable targets that we can use to *make a difference* in the disease outcomes of real patients.

This suggests that causal selection for disease is pragmatic in the sense that it is relative to the *practical* goal of control. Once the goal of control is specified, there are objective facts and considerations about what means conduce to it. These considerations include whether factors meet a minimal interventionist criterion (i.c.) in the second step of causal selection, and whether factors provide (1) specific, (2) probable, and/or (3) stable causal control in the third and final step of causal selection. These may not be the exclusive types of control

that guide this process, but they help to clarify why we often single out few factors as “the” causes of particular diseases.

## 6.0 CONCLUSION

In this project I have examined three main explanatory patterns: explanations that involve appealing to monocausal factors, causal pathways, and dynamical models. I have discussed the structure of these patterns, the rationale behind their structure, and when they are found in some contexts, while not in others. First, the monocausal case involve a situation where single factors that are viewed as the “chief” or “main” cause of some outcome of interest. This pattern is commonly used in explaining disease—for example, scientists claim that the cause of tuberculosis is the tubercle bacteria, that the cause of scurvy is a deficiency of vitamin C, and that the cause of Huntington’s disease is a mutation in the *huntingtin* gene. Second, when causal pathways are cited in explanations, they are often characterized by a (1) sequence of causal steps, where these steps (2) abstract from significant biological detail, and (3) span traditionally accepted biological levels. Examples of causal pathways include: pathways from genotype to phenotype, pathways as bundles of neurons (or neural tracts) in the spinal cord and brain, metabolic pathways, cell signaling pathways, and gene expression pathways. A third explanatory pattern involves dynamical models, which are mathematical models that describe how variables representing a particular system evolve with time. The dynamical models I have examined represent neural excitability with coupled differential equations.

Although dominant views in the philosophical literature claim that most or all of the explanations in the biological sciences are mechanistic, there are clear senses in which these patterns do not meet the mechanistic paradigm. Standard accounts define mechanisms as the underlying component parts of a system and the features, activities, and organization of these components that are relevant to the production of a particular phenomenon of interest (Machamer et al., 2000; Kaplan and Craver, 2011). In this manner, mechanistic explanation

involves appealing to the entities, activities, and organization of the relevant mechanism, where it is sometimes suggested that explanatory power increases when more of this detail is cited (Kaplan and Craver, 2011; Kaplan, 2011). However, it is a strain to interpret the monocausal model as providing a mechanistic explanation. One reason for this is that the model involves a single causal entity and mechanisms are standardly represented as having multiple interacting entities. Furthermore, scientists attribute a special explanatory status to the monocausal factor, which they distinguish from the causal relevance of other factors. This leads them to cite these single causes while abstracting from many other causal details, which they do not view as details that would increase the quality of their explanations. If providing more mechanistic detail always increases the quality of an explanation, then it isn't clear why scientists in these domains often abstract from significant amount biological detail and how they make determinations about which details are causally relevant, and which are not. This is apparent in cases of monocausal disease explanation, where scientists select and cite single factors as the most explanatorily relevant for some disease outcome of interest. In chapter 5, I discussed how this causal selection is guided, in part, by scientists' interest in identifying factors with particular types of control over the outcome of interest, i.e. control of broad scope, control that is highly probable, and control that is stable. These types of control clarify the rationale that guides scientists selection of particular causal details, and abstraction from others, in their explanations of disease phenotypes.

Many of these points also hold for pathway explanations—these explanations involve significant abstraction from detail that is unexplained by mechanistic accounts. I have suggested that features of the contexts in which pathway explanations are employed also help in clarifying the reasoning behind this explanatory pattern. One of these features has to do with the types of causal complexity that are found in these contexts. In chapter 2 I discussed two types of causal complexity that influence explanatory practice—multicausality and causal heterogeneity. When an explanatory target of interest is produced by many factors that work together in aggregate (multicausality), there is an interest in appealing to the interacting causes and understanding how they influence each other in producing the outcome of interest. Alternatively, when distinct instances of an explanatory target are caused by completely different combinations of causal factors (causal heterogeneity), appealing to any combination

provides a limited explanation, because it doesn't explain all instances of interest. I have examined one strategy for overcoming this challenge of causal heterogeneity. When a type-level phenomenon of interest has heterogeneous causes, scientists sometimes search for and identify a final common pathway that all of these factors converge on and operate through in producing the phenomenon. The final common pathway identifies factors which are shared causes across instances of the explanandum phenomenon and that "make a difference" to all of these instances ([Woodward, 2003](#)). This clarifies a method of abstracting from upstream heterogeneous factors and appealing to downstream common causes, on the basis of identifying shared causes or factors that are causally relevant to all or most of the instances of some type-level phenomenon. In chapter 2 I discussed pathway explanations in neuropsychiatry and how they are found in contexts of multicausality and causal heterogeneity. Chapter 4 examined pathway explanations in contexts of causal heterogeneity in order to suggest a type of non-reductive explanation in biology, where scientists abstract from heterogeneous lower-level causes in favor of shared (or common) causes as a higher-level (i.e. causes at a cellular level as opposed to causes at a genetic level).

An additional feature of pathway explanations that reveals one of their differences from mechanistic explanation, is that they involve causal relationships that span traditionally accepted biological levels, where causal mechanical interactions are said to be constrained to levels ([Craver, 2009a](#)). For example, in the case of Parkinson's disease, scientists cite pathways that involve causal relationships spanning genetic to cellular factors, cellular factors to neural circuitry, and neural circuitry to behavioral phenotypes. Each of these causal relationships involves factors that are often viewed as being at lower and higher biological levels, respectively. For example, Parkinson's disease is often explained by appealing to a causal pathway where (1) a gene variant results in cell death of dopaminergic neurons, (2) the death of these neurons results in neural circuitry alterations, and (3) these circuitry alterations lead to the Parkinsonian disease phenotype.<sup>1</sup> Where the notion of a causal pathway is used to refer to causal relationships that span traditionally accepted biological levels, the standard philosophical notion of a causal mechanism does not accommodate this

---

<sup>1</sup>The behavioral phenotype, in this case, is the constellation of signs and symptoms (rigidity, bradykinesia, etc.) that make-up the clinical presentation of Parkinson's disease.

feature.

A final explanatory pattern that I have examined in this project is the case of dynamical explanations in neuroscience. In this situation, scientists are interested in explaining a neural firing behavior that is exhibited across systems with differing lower-level details or differing lower-level causal mechanisms. In this case, an explanation of the shared excitatory behavior cannot involve appealing to lower-level mechanisms, because systems with this shared behavior do not have a lower-level mechanism in common. My analysis of this explanatory pattern in chapter 2 suggests that it is similar to Batterman’s discussion of minimal model explanations ([Batterman and Rice, 2014](#)). Scientists begin with mathematical models of physically distinct neural systems and reduce these models by continuously applying principled mathematical abstraction techniques. The application of these techniques reveals that all systems of interest are reduced to a common single-variable model, referred to as a canonical model. Both the canonical model and the principled mathematical abstraction techniques allow for an explanation of why these systems exhibit a shared behavior. Specifically, these systems exhibit the same behavior because they converge on the same canonical model after repeated simplification with principled techniques. Accounts of mechanistic explanation cannot accommodate this pattern, in part, because the canonical model fails to meet criteria that mechanistic philosophers claim are required of explanatory models ([Kaplan and Craver, 2011](#)). Although the canonical model fails to meet such mechanistic criteria, my analysis suggests that it plays an important role in explaining neural firing behavior.

This project suggests that explanation in biology, neuroscience, and medicine is more diverse than mechanistic accounts have suggested and that explanatory practice in these domains involves goals, techniques, and strategies that have not received significant attention in the philosophical literature. The role of these methodological considerations in explanatory practice is best understood in the context of an appreciation for: the types of explanandum scientists identify (e.g. whether it is a token or type-level phenomenon, what contrastive focus is specified, etc.), the forms of causal complexity are present for such explanatory targets (e.g. multicausality, causal heterogeneity, or other forms), and other features that they expect an explanans to have or that might serve various context-specific goals (e.g. that the explanans identifies causal factors in general or causal factors with particular types of

control). Given the variety of goals, types of causal complexity, and forms of explananda in the biological sciences, it seems unsurprising that this domain would involve explanatory patterns that are fine-tuned to these variations and distinct in important ways. A philosophical account of explanation that is relevant to biology, neuroscience, and medicine, should help clarify what these differences are, why they matter, and how they capture important features of scientific explanation.



## REFERENCES

- Abbott, L. F. (1994). Single Neuron Dynamics: An Introduction. In F. Ventriglia (Ed.), *Neural Modeling and Neural Networks*, pp. 57–78. Pergamon Press.
- Achinstein, P. (1984). The Pragmatic Character of Explanation. pp. 1–19. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.
- Barabási, A.-L. and Z. N. Oltvai (2004, February). Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics* 5(2), 101–113.
- Batterman, R. Problems with Scientific Reductionism (Forthcoming). pp. 1–23. Batterman, R. (2000). A ‘Modern’ (= Victorian?) Attitude Towards Scientific Understanding. *The Monist* 83, 228–257.
- Batterman, R. and C. Rice (2014). Minimal Model Explanations. *Philosophy of Science* 81.
- Batterman, R. W. (2001). *The Devil in the Details*. Asymptotic Reasoning in Explanation, Reduction, and Emergence. Oxford University Press, USA.
- Batterman, R. W. (2002). Asymptotics and the Role of Minimal Models. *The British Journal for the Philosophy of Science* 53(1), 21–38.
- Batterman, R. W. (2010, February). On the Explanatory Role of Mathematics in Empirical Science. *The British Journal for the Philosophy of Science* 61(1), 1–25.
- Bean, B. P. (2007). The action potential in mammalian central neurons. *Nature Reviews Neuroscience* 8(6), 451–465.
- Bechtel, W. and J. Mundale (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 175–207.
- Bickle, J. (2006). Reducing Mind to Molecular Pathways: Explicating the Reductionism Implicit in Current Cellular and Molecular Neuroscience. *Synthese* 151(3), 411–434.

- Blaxter, M. (2015). *Health*. Polity Press.
- Boorse, C. (1977). Health as a Theoretical Concept. *Philosophy of Science*, 1–33.
- Börger, C., S. Epstein, and N. J. Kopell (2008). Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proceedings of the National Academy of Sciences* 105(46), 18023–18028.
- Brady, S. T., G. J. Siegel, R. W. Albers, and D. L. Price (2012). *Basic Neurochemistry* (8 ed.). Principles of Molecular, Cellular, and Medical Neurobiology. Elsevier.
- Bryant, P. A., J. Trinder, and N. Curtis (2004). Sick and Tired: Does Sleep Have a Vital Role in the Immune System? *Nature Reviews Immunology* 4(6), 457–467.
- Burbulla, L. F. and R. Krüger (2011). Converging environmental and genetic pathways in the pathogenesis of Parkinson’s disease. *Journal of the Neurological Sciences* 306(1-2), 1–8.
- Butterfield, J. (2011). Emergence, Reduction and Supervenience: A Varied Landscape. *Foundations of Physics* 41(6), 920–959.
- Cannon, T. D. and M. C. Keller (2006). Endophenotypes in the Genetic Analyses of Mental Disorders. *Annual Review of Clinical Psychology* 2(1), 267–290.
- Caspi, A. and T. E. Moffitt (2006). Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews Neuroscience* 7, 1–8.
- Cauli, B., E. Audinat, B. Lambolez, M. C. Angulo, N. Ropert, K. Tsuzuki, S. Hestrin, and J. Rossier (1997). Molecular and physiological diversity of cortical nonpyramidal cells. *The Journal of Neuroscience* 17(10), 3894–3906.
- Chirimuuta, M. (2013). Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience. *Synthese*.
- Connor, J. A. (1975). Neural repetitive firing: a comparative study of membrane properties of crustacean walking leg axons. *Journal of Neurophysiology* 38(4), 922–932.
- Corti, O., S. Lesage, and A. Brice (2011). What Genetics Tells us About the Causes and Mechanisms of Parkinson’s Disease. *Physiological Reviews* 91(4), 1161–1218.
- Craver, C. and J. Tabery (2015). Mechanisms in Science.
- Craver, C. F. (2006). When mechanistic models explain. *Synthese* 153(3), 355–376.
- Craver, C. F. (2008). Physical Law and Mechanistic Explanation in the Hodgkin and Huxley

- Model of the Action Potential. *Philosophy of Science* 75(5), 1022–1033.
- Craver, C. F. (2009a). *Explaining the Brain*. OUP Oxford.
- Craver, C. F. (2009b). Mechanisms and natural kinds. *Philosophical Psychology* 22(5), 575–594.
- Craver, C. F. and W. Bechtel (2007). Top-down causation without top-down causes. *Biology & Philosophy* 22(4), 547–563.
- Dauer, W. and S. Przedborski (2003). Parkinson’s Disease: Mechanisms and Models. *Neuron*, 889–909.
- Doi, S. and S. Kumagai (2001). Nonlinear Dynamics of Small-Scale Biophysical Neural Networks. In R. R. Poznanski (Ed.), *Biophysical Neural Networks: Foundations of Integrative Neuroscience*, pp. 261–302. Mary Ann Liebert, Inc. Publishers.
- Doweiko, H. (1999). *Are People Predestined to Become Addicted to Chemicals?* Concepts of Chemical Dependency. Brooks/Cole Publishing Company.
- Dupré, J. (2013). Living Causes. *Aristotelian Society Supplementary Volume* 87(1), 19–37.
- Ermentrout, B., J. Rubin, and R. Osan (2002). Regular traveling waves in a one-dimensional network of theta neurons. *SIAM Journal on Applied Mathematics* 62(4), 1197–1221.
- Ermentrout, G. B. and D. H. Terman (2010). *Mathematical Foundations of Neuroscience*, Volume 35 of *Interdisciplinary Applied Mathematics*. Springer.
- Fitzhugh, R. (1960). Thresholds and plateaus in the Hodgkin-Huxley nerve equations. *The Journal of General Physiology* 43(5), 867–896.
- Fitzhugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal* 1(6), 445–466.
- Fodor, J. (1975). *The Language of Thought*. Thomas Crowell.
- Fodor, J. (1997). Special Sciences: Still Authonomous After All These Years. *Nous*, 1–16.
- Fodor, J. A. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, 1–20.
- Fowler, A. C. (2007). *Mathematical Models in the Applied Sciences*. Cambridge University Press.
- Geschwind, D. H. (2008). Autism: Many Genes, Common Pathways? *Cell* 135(3), 391–395.
- Geschwind, D. H. (2011). Genetics of autism spectrum disorders. *Trends in Cognitive*

- Sciences* 15(9), 409–416.
- Goldenfeld, N., O. Martin, and Y. Oono (1989). Intermediate Asymptotics and Renormalization Group Theory . *Scientific Computing* 4, 1–19.
- Gottesman, I. I. and J. Shields (1972). *Schizophrenia and Genetics: as Twin Study Vantage Point* . Academic Press.
- Griffiths, P. E. Proximate and Ultimate Information in Biology (Formthing). In *Festschrift for Philip Kitcher*. Oxford University Press.
- Gutkin, B. S. and B. G. Ermentrout (1998). Dynamics of Membrane Excitability Determine Interspike Interval Variability: A Link Between Spike Generation Mechanisms and Cortical Spike Train Statistics. *Neural computation*, 1047–1065.
- Hecht, S. S. (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature Reviews Cancer* 3(10), 733–744.
- Hempel, C. (1965). *Aspects of Scientific Explanation*. And Other Essays in The Philosophy of Science. The Free Press.
- Hitchcock, C. and J. Woodward (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Nous* 37(2), 181–199.
- Hodgkin, A. L. (1948). The Local Electric Changes Associated with Repetitive Action in Non-Medullated Axon. *Journal of Physiology* (107), 165–181.
- Hoppensteadt, F. C. and E. M. Izhikevich (1997). *Weakly Connected Neural Networks*. Springer-Verlag New York Incorporated.
- Izhikevich, E. M. (2004). Which Model to Use for Cortical Spiking Neurons? *IEEE Transactions on Neural Networks* 15(5), 1063–1070.
- Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience*. MIT Press (MA).
- Jia, B., H.-G. Gu, and Y.-Y. Li (2011). Coherence-Resonance-Induced Neuronal Firing near a Saddle-Node and Homoclinic Bifurcation Corresponding to Type-I Excitability. *Chinese Physics Letters* 28(9), 090507.
- John, B. and K. R. Lewis (1966). Chromosome variability and geographic distribution in insects. *Science* 152, 711–721.
- Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese* 183(3), 339–373.

- Kaplan, D. M. and C. F. Craver (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science* 78(4), 601–627.
- Kendler, K. S. (2005). Psychiatric Genetics: A Methodological Critique. *American Journal of Psychiatry*, 3–11.
- Kendler, K. S. (2011). Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology. *Molecular Psychiatry* 17(1), 11–21.
- Kendler, K. S. (2013). What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. *Molecular Psychiatry* 18(10), 1058–1066.
- Kendler, K. S. and M. C. Neale (2010). Endophenotype: a conceptual analysis. *Molecular Psychiatry* 15(8), 789–797.
- Kendler, K. S., P. Zachar, and C. Craver (2010). What kinds of things are psychiatric disorders? *Psychological Medicine* 41(06), 1143–1150.
- Kennan, A., A. Aherne, and P. Humphries (2005). Light in retinitis pigmentosa. *Trends in Genetics* 21(2), 103–110.
- Kiesler, D. (1999). Beyond the Disease Model of Mental Disorders. *Greenwood Publishing Group*, 1–244.
- Kincaid, H. and J. Sullivan (2014). Classifying Psychopathology: Mental Kinds and Natural Kinds. In H. Kincaid and J. Sullivan (Eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*. The MIT Press.
- Kitcher, P. (1984). 1953 and all that. A tale of two sciences. *The Philosophical Review*, 335–373.
- Klein, C. and M. G. Schlossmacher (2006). The genetics of Parkinson disease: implications for neurological care. *Nature Clinical Practice Neurology* 2(3), 136–146.
- Lesage, S. and A. Brice (2009). Parkinson’s disease: from monogenic forms to genetic susceptibility factors. *Human Molecular Genetics* 18(R1), R48–R59.
- Lewis, D. (1973). Causation. *The Journal of Philosophy*.
- Lewis, D. (1986). Causal Explanation. In *Philosophical Papers*. Oxford University Press.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mecha-

- nisms influence causal ascriptions. *Cognitive Psychology* 61(4), 303–332.
- Lu, H., A. L. Yuille, M. Liljeholm, P. W. Cheng, and K. J. Holyoak (2008). Bayesian generic priors for causal learning. *Psychological Review* 115(4), 955–984.
- Lu, L. J., A. Sboner, Y. J. Huang, H. X. Lu, T. A. Gianoulis, K. Y. Yip, P. M. Kim, G. T. Montelione, and M. B. Gerstein (2007). Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends in Biochemical Sciences* 32(7), 320–331.
- Luo, X., L. Huang, P. Jia, M. Li, B. Su, Z. Zhao, and L. Gan (2014). Protein-Protein Interaction and Pathway Analyses of Top Schizophrenia Genes Reveal Schizophrenia Susceptibility Genes Converge on Common Molecular Networks and Enrichment of Nucleosome (Chromatin) Assembly Genes in Schizophrenia Susceptibility Loci. *Schizophrenia Bulletin* 40(1), 39–49.
- Machamer, P. and J. Bogen (2011). Mechanistic information and causal continuity. In P. M. Illari, F. Russo, and J. Williamson (Eds.), *Causality in the Sciences*. Oxford Scholarship Online.
- Machamer, P., L. Darden, and C. F. Craver (2000). Thinking about mechanisms. *Philosophy of Science* 67, 1–25.
- Martins-de Souza, D. (2012). Proteomics Tackling Schizophrenia as a Pathway Disorder. *Schizophrenia Bulletin* 38(6), 1107–1108.
- McClellan, J. and M.-C. King (2010a). Genetic Heterogeneity in Human Disease. *Cell* 141(2), 210–217.
- McClellan, J. and M.-C. King (2010b). Genomic Analysis of Mental Illness. *JAMA: The Journal of the American Medical Association*, 2523–2524.
- Mill, J. S. (1874). *A System of Logic* (Eighth ed.). Harper & Brothers Publishers.
- Murphy, D. (2006). The Medical Model and the Foundations of Psychiatric Explanation. In *Psychiatry in the Scientific Image*, pp. 1–46. The MIT Press.
- Murphy, D. (2014). Natural Kinds in Folk Psychology and in Psychiatry. In H. Kincaid and J. Sullivan (Eds.), *Classifying psychopathology : mental kinds and natural kinds*. The MIT Press.
- Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simu-

- lating nerve axon. *Proceedings of the IRE* 50(10), 2061–2070.
- Nestler, E. J. (2005, November). Is there a common molecular pathway for addiction? *Nature Neuroscience* 8(11), 1445–1449.
- Obeso, J. A., M. C. Rodriguez-Oroz, C. G. Goetz, C. Marin, J. H. Kordower, M. Rodriguez, E. C. Hirsch, M. Farrer, A. H. V. Schapira, and G. Halliday (2010). Missing pieces in the Parkinson’s disease puzzle. *Nature medicine* 16(6), 653–661.
- Papin, J. A., N. D. Price, S. J. Wiback, D. A. Fell, and B. O. Palsson (2003, May). Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences* 28(5), 250–258.
- Plomin, R., J. C. DeFries, V. S. Knopick, and J. M. Neiderhiser (2012). Pathways between Genes and Behavior. In *Behavioral Genetics*, pp. 1–10. Worth Publishers.
- Poland, J. S. (2014). Deeply Rooted Sources of Error and Bias in Psychiatric Classification. In H. Kincaid and J. Sullivan (Eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*. The MIT Press.
- Putnam, H. (1975). *Philosophy and our Mental Life*. Mind Language and Reality. Cambridge University Press.
- Rinzel, J. and G. B. Ermentrout (1989). Analysis of neural excitability and oscillations. In *Methods in Neuronal Modelling: From synapses to Networks*, pp. 135–169. Cambridge, MA: MIT Press.
- Schaffer, J. (2014). The Metaphysics of Causation.
- Schaffner, K. F. (1993). *Discovery and Explanation in Biology and Medicine*. University Of Chicago Press.
- Schaffner, K. F. (1998). Genes, Behavior, and Developmental Emergentism: One Process, Indivisible? *Philosophy of Science* 65, 209–252.
- Schaffner, K. F. (2008). *Etiological Models in Psychiatry: Reductive and Nonreductive Approaches*. Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology. Johns Hopkins Press.
- Schaffner, K. F. (2012). A philosophical overview of the problems of validity for psychiatric disorders. In K. Kendler and J. Parnas (Eds.), *Philosophical Issues in Psychiatry II*, pp. 1–32. Oxford University Press.
- Scriven, M. (1959). Explanation and Prediction in Evolutionary Theory. *Science*, 1–7.

- Shulman, J. M., P. L. De Jager, and M. B. Feany (2011). Parkinson’s Disease: Genetics and Pathogenesis. *Annual Review of Pathology: Mechanisms of Disease* 6(1), 193–222.
- Silberstein, M. and A. Chemero (2008). Replacing Scholasticism with Science. *Philosophy of Science* 75(1), 1–27.
- Silberstein, M. and A. Chemero (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science* 80(5), 958–970.
- Snustad, D. P. and M. J. Simmons (2012). *Principles of Genetics*. Wiley.
- Sober, E. (1999). The Multiple Realizability Argument against Reductionism. *Philosophy of Science*, 1–24.
- Stein, D. J. (2014). Psychopharmacology and Natural Kinds: A Conceptual Framework. In H. Kincaid and J. Sullivan (Eds.), *Classifying psychopathology: mental kinds and natural kinds*. The MIT Press.
- Stepp, N., A. Chemero, and M. T. Turvey (2011). Philosophy for the Rest of Cognitive Science. *Topics in Cognitive Science* 3(2), 425–437.
- Sullivan, P. F. (2012). Schizophrenia as a pathway disease. *Nature medicine*, 1–2.
- Sullivan, P. F., M. J. Daly, and M. O’Donovan (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Publishing Group* 13(8), 537–551.
- Tateno, T. (2004). Threshold Firing Frequency-Current Relationships of Neurons in Rat Somatosensory Cortex: Type 1 and Type 2 Dynamics. *Journal of Neurophysiology* 92(4), 2283–2294.
- Vacher, H., D. P. Mohapatra, and J. S. Trimmer (2008). Localization and Targeting of Voltage-Dependent Ion Channels in Mammalian Central Neurons. *Physiological Reviews* 88(4), 1407–1447.
- Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics* 90(1), 7–24.
- Waskan, J. (2011, January). Mechanistic explanation at the limit. *Synthese* 183(3), 389–408.
- Waters, C. K. (1990). Why the Anti-Reductionist Consensus Won’t Survive: The Case of Classical Mendelian Genetics. *PSA Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1–16.



- Waters, C. K. (2007). Causes That Make a Difference. *The Journal of Philosophy*, 1–30.
- Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings*. Harvard University Press.
- Woodward, J. The Problem of Variable Choice (Forthcoming). *Synthese*, 1–29.
- Woodward, J. (2003). *Making Things Happen*. Oxford University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review* 115(1), 1–50.
- Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy* 25(3), 287–318.
- Woodward, J. (2013, June). II-Mechanistic Explanation: Its Scope and Limits. *Aristotelian Society Supplementary Volume* 87(1), 39–65.
- Woodward, J. and C. Hitchcock (2003). Explanatory generalizations, part I: A counterfactual account. *Nous* 37(1), 1–24.
- Yang, X. (2009). Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nature Genetics* 41(4), 415–423.
- Zachar, P. (2014). Beyond Natural Kinds: Toward a "Relevant" "Scientific" Taxonomy in Psychiatry. In H. Kincaid and J. Sullivan (Eds.), *Classifying psychopathology: mental kinds and natural kinds*. The MIT Press.